Optimization over digraphs: Linear algorithms with linear convergence

A thesis submitted by

Ran Xin

in partial fulfillment of the requirements for the degree of

Master of Science

in

Electrical Engineering

Tufts University

Advisor: Professor Usman Ahmed Khan

May 2018

Declaration of Authorship

I, Ran Xin, declare that this thesis titled, "Optimization over digraphs: Linear algorithms with linear convergence" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Ran Xin

Date: April 30, 2018

"The essence of mathematics is not to make simple things complicated, but to make complicated things simple."

Stanley Gudder

TUFTS UNIVERSITY

Abstract

Electrical Engineering

Master of Science

Optimization over digraphs: Linear algorithms with linear convergence

by Ran Xin

In this thesis, we study distributed optimization, where a network of agents, interacting over a directed graph, collaborates to minimize the average of locally-known convex functions. Most of the existing algorithms apply push-sum consensus, which utilizes column-stochastic weight matrices. Column-stochastic weights require each agent to know (at least) its out degree, which may be impractical in e.g., broadcast-based communication protocols. In contrast, we describe FROST (Fast Row-stochastic Optimization with uncoordinated STep-sizes), an optimization algorithm applicable to directed graphs with row-stochastic weights and non-identical step-sizes at the agents. Its implementation is straightforward as each agent locally decides the weights assigned to the incoming information and locally chooses a suitable step-size. Furthermore, we propose a completely linear algorithm which avoids using push-sum (type) techniques and thus leads to less communication and computation over the network of agents. Under the assumptions that each local function is strongly-convex with Lipschitz-continuous gradients, we show that the proposed algorithms linearly converge to the global minimizer with sufficiently small step-sizes. We present numerical simulations to illustrate our theoretical results.

Acknowledgements

I gratefully thank my advisor, Professor Usman Khan, for his guidance and encouragement through my research in the last two years. I also deeply thank my thesis committe members, Prof. Brian Tracey from Electrical Engineering department, Tufts University, Prof. Babak Moaveni from Civil and Environmental Engineering Department, Tufts University, for their invaluable discussions and suggestions.

Contents

Dedicated to my parents

Chapter 1

Introduction

In this thesis, we consider distributed optimization over directed multi-agent networks. Formally, each agent i has access only to a private function, $f_i : \mathbb{R}^p \to \mathbb{R}$. The goal is to minimize the average of these functions, $\frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$, via information exchange among the agents. We focus on the case where the communication network is described by an arbitrary *directed* graph. Some initial methods for distributed optimization include distributed sub-gradient descent (DGD) Nedić and Ozdaglar, 2009, which converges to the optimal solution at a sublinear rate, i.e., $O(\frac{\ln k}{\sqrt{k}})$ for arbitrary (possibly non-differentiable) convex functions and $O(\frac{\ln k}{k})$ for strongly-convex functions, where k is the number of iterations. These methods are slow due to the diminishing step-sizes. With the help of strong-convexity and Liptschiz-continuous gradients, algorithms with faster convergence rates have been developed. In particular, DGD with constant step-size Yuan, Ling, and Yin, 2016 is shown to have a linear convergence rate to an error ball around the optimal solution. Another method, EXTRA Shi et al., 2015, achieves linear convergence to the global optimal with the requirement of symmetric weights. Of relevance are Refs. Xu et al., 2015; Qu and Li, 2017b; Xu et al., 2018; Qu and Li, 2017a, which combine inexact gradient methods and a gradient estimation technique based on dynamic consensus Zhu and Martínez, 2010. Additional related work and applications can be found in Mota et al., 2012; Jakovetic, 2017; Raja and Bajwa, 2016; Lee and Zavlanos, 2017; Mansoori and Wei, 2017; Ying and Sayed, 2018.

All the aforementioned methods require the underlying graphs to be undirected or weight-balanced. This requirement, however, may not be practical, for example, when the agents broadcast at different power levels. It is natural thus to develop

optimization and learning algorithms that are applicable to directed graphs. The primary challenge in dealing with directed graphs is that it may not be possible to construct doubly-stochastic weight matrices for information fusion. The weighted adjacency matrix for directed graphs, in general, may only be either row-stochastic or column-stochastic, but not both. See Gharesifard and Cortés, 2012 for work on balancing the weights in strongly-connected directed graphs.

The existing approaches for optimization over directed graphs are motivated by combining average-consensus methods developed for directed graphs with optimization algorithms designed for undirected graphs. For instance, subgradient-push introduced in Tsianos, Lawlor, and Rabbat, 2012b and further studied in Nedić and Olshevsky, 2015 combines push-sum consensus Kempe, Dobra, and Gehrke, 2003 and DGD; A linear algorithm over directed graphs, called Directed-Distributed Gradient Descent (D-DGD), was introduced in Xi, Wu, and Khan, 2017; Xi and Khan, 2016, and is based on surplus consensus Cai and Ishii, 2012 and DGD. Such DGD-based methods, however, restricted by the diminishing step-size, converge relatively slowly at $O(\frac{\ln k}{\sqrt{k}})$ for general convex functions and $O(\frac{\ln k}{k})$ for strongly-convex functions. The convergence rate has been recently improved in DEXTRA Xi and Khan, 2017a, which converges linearly to the global optimal given that its step-size lies in an interval and the objective functions are strongly-convex with Lipschitz-continuous gradients. DEXTRA was subsequently improved in ADD-OPT/Push-DIGing Xi, Xin, and Khan, 2017a; Nedić, Olshevsky, and Shi, 2017, which linearly converges with a sufficiently small step-size. The implementation of DEXTRA and ADD-OPT/Push-DIGing requires each agent to know its out-degree in order to construct a column-stochastic weight matrix. This requirement is impractical in many situations, especially when the agents use a broadcast-based communication protocol. In contrast to previous methods, Ref. Xi et al., 2018 provides a fast algorithm that uses only row-stochastic weights. The advantage of row-stochastic weights is the simplicity in its implementation and applicability to a broad range of communication protocols including the broadcast-based methods. These advantages, however, come at a price as the corresponding algorithm requires unique identifiers at each agent.

In this thesis, we first focus on row-stochastic weights following the recent work

in Xi et al., 2018 and describe an algorithm that we call FROST (Fast Row-stochastic Optimization with uncoordinated STep-sizes). Next, we propose a completely *linear* algorithm which avoids using push-sum (type) techniques and thus leads to less communication and computation over the network of agents.

Notation: We use lowercase bold letters to denote vectors and uppercase italic letters to denote matrices. The matrix, I_n, represents the $n \times n$ identity, whereas $\mathbf{1}_n$ is the n-dimensional column vector of all 1's. For an arbitrary vector, $\mathbf{x}$, we denote its ith element by $[\mathbf{x}]_i$ and $\text{diag}\{\mathbf{x}\}$ is a diagonal matrix with ith element on its diagonal being $[\mathbf{x}]_i$. We denote by $X \otimes Y$, the Kronecker product of two matrices, X and Y. For any $f(\mathbf{x})$, $\nabla f(\mathbf{x})$ denotes the gradient of f at $\mathbf{x}$. The spectral radius of a matrix, X, is represented by $\rho(X)$. For a primitive, row-stochastic matrix, $\underline{A}$, we denote its left and right eigenvectors corresponding to the eigenvalue of 1 by $\boldsymbol{\pi}_r$ and $\mathbf{1}_n$, respectively, such that $\boldsymbol{\pi}_r^\top \mathbf{1}_n = 1$. Similarly, for a primitive, column-stochastic matrix, $\underline{B}$, we denote its left and right eigenvectors corresponding to the eigenvalue of 1 by $\mathbf{1}_n$ and $\boldsymbol{\pi}_c$, respectively, such that $\mathbf{1}_n^\top \boldsymbol{\pi}_c = 1$. For a matrix X, we denote X_∞ as its infinite power (if it exists), i.e., $X_\infty = \lim_{k\to\infty} X^k$. We use $|||\cdot|||$ to denote matrix norms and $\|\cdot\|$ to denote vector norms. For a matrix norm, $|||\cdot|||$, there exists a compatible vector norm, $\|\cdot\|$, such that $\|X\mathbf{x}\| \leq |||X|||\|\mathbf{x}\|$, for all matrices, X, and all vectors, $\mathbf{x}$, see Theorem 5.7.13 in Horn and Johnson, 2013. The notation $\|\cdot\|_2$ denotes the Euclidean norm of vectors and $|||\cdot|||_2$ denotes the spectral norm of matrices, while $\|\cdot\|_F$ denotes the Frobenius norm of matrices.

Chapter 2

Problem formulation and related work

2.1 Problem Formulation

Consider a strongly-connected network of n agents communicating over a directed graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of agents and $\mathcal{E}$ is the set of edges, $(i, j), i, j \in \mathcal{V}$, such that agent j can send information to agent i, $j \to i$. Define $\mathcal{N}_i^{\text{in}}$ as the collection of in-neighbors, i.e., the set of agents that can send information to agent i. Similarly, $\mathcal{N}_i^{\text{out}}$ is the set of out-neighbors of agent i. Note that both $\mathcal{N}_i^{\text{in}}$ and $\mathcal{N}_i^{\text{out}}$ include node i. We focus on solving the following distributed optimization problem over the above network:

$$\text{P1}: \quad \min F(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}),$$

where each objective, $f_i : \mathbb{R}^p \to \mathbb{R}$, is a private cost function known only by agent i. We discuss this problem under different set of assumptions.

Assumption 1. *The network topology and objective functions satisfy the following:*

1. *The underlying communication graph is strongly-connected.*

2. *Each local objective function, f_i, is convex.*

Assumption 2. *Each local function, f_i, is differentiable and strongly-convex, and has globally Lipschitz-continuous gradients, i.e., for any i and $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^p$,*

1. *there exists a positive constant β such that*

$$\|\nabla f_i(\mathbf{x}_1) - \nabla f_i(\mathbf{x}_2)\|_2 \leq \beta \|\mathbf{x}_1 - \mathbf{x}_2\|_2; \tag{2.1}$$

2. *there exists a positive constant α such that,*

$$f_i(\mathbf{x}_1) - f_i(\mathbf{x}_2) \leq \nabla f_i(\mathbf{x}_1)^\top (\mathbf{x}_1 - \mathbf{x}_2) - \frac{\alpha}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2. \tag{2.2}$$

Clearly, the Lipschitz-continuity and strongly-convexity constants for the global objective function $F(\mathbf{x})$ are β and α, respectively.

Assumption 3. *Each function, f_i, has bounded subgradients.*

Assumption 4. *The agents have and know their unique identifiers, e.g., $1, \cdots, n$.*

Assumption 5. *Each agent in the network knows its out-degree.*

We note here that Assumptions 2 and 3 do not hold together. When applicable, the algorithms we discuss use either one of these assumptions but not both.

2.2 Related work

In this section, we discuss algorithms that are related to the Problem P1 and provide an intuitive explanation of each one of them.

2.2.1 Algorithms using doubly-stochastic Weights

A well-known solution to distributed optimization over undirected graphs is Distributed Gradient Descent (DGD) Nedić and Ozdaglar, 2009, which is the combination of average-consensus and a local gradient descent step. At each agent, DGD implements the following iterations:

$$\mathbf{x}_i(k+1) = \sum_{j=1}^{n} w_{ij}\mathbf{x}_j(k) - \eta(k)\nabla f_i(\mathbf{x}_i(k)), \tag{2.3}$$

where $\mathbf{x}_i(k)$ is the estimate of the optimal solution, $\mathbf{x}^*$, at agent i, k is the discrete time index, $W = \{w_{ij}\}$ is a doubly-stochastic, weighted adjacency matrix of the

communication graph. The variable $\eta(k)$ is a diminishing step-size, which satisfies the persistence condition: $\sum_{k=0}^{\infty}\eta(k) = \infty$ and $\sum_{k=0}^{\infty}\eta(k)^2 < \infty$. Under the Assumptions 1 and 3, DGD converges to the global optimal of Problem P1 with the convergence rate of $O(\frac{\ln k}{\sqrt{k}})$. The convergence rate is slow because of the diminishing step-size. If constant step-size is used in DGD, i.e., $\eta(k) = \eta$, DGD converges faster to an error ball around the optimal solution, Yuan, Ling, and Yin, 2016. This is because $\sum_{i=1}^{n}\nabla f_i(\mathbf{x}^*) = 0$ does not necessarily mean $\nabla f_i(\mathbf{x}^*) = 0$, for any i.

To achieve a fast convergence and that to the exact optimal solution of P1, Refs. Qu and Li, 2017b; Xu et al., 2015 propose a class of distributed algorithms, based on distributed inexact gradient method and dynamic average consensus, Zhu and Martínez, 2010. It can be summarized as having the following form:

$$\mathbf{x}_i(k+1) = \sum_{j=1}^{n} w_{ij}\mathbf{x}_j(k) - \eta\mathbf{y}_i(k), \tag{2.4a}$$

$$\mathbf{y}_i(k+1) = \sum_{j=1}^{n} w_{ij}\mathbf{y}_j(k) + \nabla f_i\big(\mathbf{x}_i(k+1)\big) - \nabla f_i\big(\mathbf{x}_i(k)\big), \tag{2.4b}$$

initialized with $\mathbf{y}_i(0) = \nabla f_i(\mathbf{x}_i(0))$ and arbitrary $\mathbf{x}_i(0), \forall i$. The first equation is a distributed inexact gradient method where the local descent direction is $\mathbf{y}_i(k)$ instead of $\nabla f_i(\mathbf{x}_i(k))$ as was in Eq. (2.3). The second equation is a gradient estimation technique when viewed as dynamic consensus Zhu and Martínez, 2010, where $\mathbf{y}_i(k)$ tracks the average of local gradients $\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i(k))$. It is shown in Ref. Qu and Li, 2017b that Eqs. (2.4) converge linearly to the optimal solution of Problem P1 under Assumptions 1 and 2 as long as the step-size, η, is sufficiently small. Note that these methods, Eq. (2.3) and Eqs. (2.4), are not applicable to directed graphs since the required doubly-stochastic weight matrix cannot be constructed in arbitrary directed graphs.

2.2.2 Algorithm using column-stochastic weights

We now consider the case when DGD is applied to a directed graph, where the weight matrix, W, is chosen to be column-stochastic implying that it cannot be row-stochastic at the same time. Denote $\overline{\mathbf{x}}(k) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i(k)$ as the average of the estimates at all agents at iteration k. It can be obtained that Xi, Wu, and Khan, 2017:

$$\overline{\mathbf{x}}(k+1) = \overline{\mathbf{x}}(k) - \frac{\eta(k)}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i(k)). \tag{2.5}$$

From Eq. (2.5), it is clear that the average of the estimates at each agent converges to the global optimal solution of Problem P1 since Eq. (2.5) can be viewed as a centralized gradient method with the correct gradient direction if the local estimate $\mathbf{x}_i(k)$ converges to $\overline{\mathbf{x}}(k)$. However, since the weight matrix is *not row-stochastic*, a necessary condition for agreement over graphs, the agent estimates will not reach an agreement, see e.g., Xi, Wu, and Khan, 2017 for additional details. This discussion motivates the idea of combining DGD with an algorithm, called push-sum briefly discussed next, that enables agreement over directed graphs with column-stochastic weights.

Push-sum algorithm

Push-sum Benezit et al., 2010; Kempe, Dobra, and Gehrke, 2003 is a technique for a network of agents to achieve average-consensus over arbitrary directed graphs. At kth iteration, each agent maintains two state vectors, $\mathbf{x}_i(k), \mathbf{z}_i(k) \in \mathbb{R}^p$, and an auxiliary scalar variable, $y_i(k)$, initialized with $y_i(0) = 1$. Push-sum performs the following iterations:

$$\mathbf{x}_i(k+1) = \sum_{j=1}^{n} b_{ij}\mathbf{x}_j(k) \tag{2.6a}$$

$$y_i(k+1) = \sum_{j=1}^{n} b_{ij}y_j(k), \tag{2.6b}$$

$$\mathbf{z}_i(k+1) = \frac{\mathbf{x}_i(k+1))}{y_i(k+1)}, \tag{2.6c}$$

where $\underline{B} = \{b_{ij}\}$ is a primitive and column-stochastic weight matrix. Eq. (2.6b) can be viewed as an independent algorithm to asymptotically learn the right eigenvector

of $\underline{B}$; recall that the right eigenvector of $\underline{B}$, corresponding to the eigenvalue 1, is not $\mathbf{1}_n$ because $\underline{B}$ is not row-stochastic and we denote this right eigenvector by $\boldsymbol{\pi}_c$. In fact, we can show that $\lim_{k\to\infty} y_i(k) = n[\boldsymbol{\pi}_c]_i$. Besides, it can be verified that $\lim_{k\to\infty} \mathbf{x}_i(k) = [\boldsymbol{\pi}_c]_i \sum_{i=1}^n \mathbf{x}_i(0)$. Therefore, the limit of $\mathbf{z}_i(k)$, as the ratio of $\mathbf{x}_i(k)$ over $y_i(k)$, achieves the agreement at the average of the initial values, i.e.,

$$\lim_{k\to\infty} \mathbf{z}_i(k) = \lim_{k\to\infty} \frac{\mathbf{x}_i(k)}{y_i(k)} = \frac{[\boldsymbol{\pi}_c]_i \sum_{i=1}^n \mathbf{x}_i(0)}{n[\boldsymbol{\pi}_c]_i} = \frac{\sum_{i=1}^n \mathbf{x}_i(0)}{n}.$$

In the next subsection, we present the subgradient-push algorithm which applies push-sum consensus to DGD, see Xi, Wu, and Khan, 2017 for an alternate linear approach.

Subgradient-Push

To solve Problem P1 over arbitrary directed graphs, Refs. Tsianos, Lawlor, and Rabbat, 2012b; Tsianos, 2013; Tsianos, Lawlor, and Rabbat, 2012a; Nedić and Olshevsky, 2015 develop subgradient-push, which performs the following iterations:

$$\mathbf{x}_i(k+1) = \sum_{j=1}^n b_{ij}\mathbf{x}_j(k) - \eta(k)\nabla f_i\big(\mathbf{z}_i(k)\big), \tag{2.7a}$$

$$y_i(k+1) = \sum_{j=1}^n b_{ij} y_j(k), \tag{2.7b}$$

$$\mathbf{z}_i(k+1) = \frac{\mathbf{x}_i(k+1))}{y_i(k+1)}. \tag{2.7c}$$

Subgradient-Push is initialized at $y_i(0) = 1, \forall i$, and with arbitrarily chosen $\mathbf{x}_i(0)$'s. The step-size, $\eta(k)$, satisfies the same persistence conditions as DGD. To understand these iterations, note that Eqs.(2.7) are nearly the same as Eqs. (2.6) except that there is a gradient term in Eq. (2.7a), which drives the limit of $\mathbf{z}_i(k)$ to the global optimal solution of P1. Under the Assumptions 1, 3 and 5, subgradient-push converges at the rate of $O(\frac{\ln k}{\sqrt{k}})$. We next provide an algorithm that significantly improves this convergence rate.

ADD-OPT

ADD-OPT Xi, Xin, and Khan, 2017b is a fast algorithm over directed graphs, which converges at a linear rate to the global optimal solution under the Assumptions 1, 2, and 5, in contrast to the sublinear convergence of subgradient-push. The three vectors, $\mathbf{x}_i(k)$, $\mathbf{z}_i(k)$, $\mathbf{w}_i(k)$, as well as a scalar $y_i(k)$ maintained at each agent i, are updated as follows:

$$\mathbf{x}_i(k+1) = \sum_{j=1}^{n} b_{ij}\mathbf{x}_j(k) - \eta \mathbf{w}_i(k), \tag{2.8a}$$

$$y_i(k+1) = \sum_{j=1}^{n} b_{ij} y_j(k), \tag{2.8b}$$

$$\mathbf{z}_i(k+1) = \frac{\mathbf{x}_i(k+1))}{y_i(k+1)}. \tag{2.8c}$$

$$\mathbf{w}_i(k+1) = \sum_{j=1}^{n} b_{ij}\mathbf{w}_j(k) + \nabla f_i\big(\mathbf{z}_i(k+1)\big) - \nabla f_i(\mathbf{z}_i(k)), \tag{2.8d}$$

where $\underline{B} = \{b_{ij}\}$ is column-stochastic, $y_i(0) = 1$, $\mathbf{w}_i(0) = \nabla f_i(\mathbf{x}_i(0)), \forall i$, and $\mathbf{x}_i(0)$'s are arbitrary. We note here that ADD-OPT essentially applies push-sum to the algorithm in Eqs. (2.4). And Eq.(2.8d) estimates the average of local gradients with column-stochastic weights.

All the aforementioned methods over directed graphs require each agent to know its out-degree in order to construct the column-stochastic weight matrices. This requirement may be infeasible especially when agents use broadcast-based communication protocols. Row-stochastic weights, on the other hand, are much easier to implement in a distributed manner as each agent locally decides the weights assigned to each incoming variable from the neighboring agents.

2.2.3 Algorithms using only Row-stochastic Weights

Lets first consider DGD in Eq. (2.3) when the weight matrix W is only row-stochastic. From average-consensus Xi, Wu, and Khan, 2017 and the fact that the step-size $\eta(k)$ goes to 0, it can be verified that the agents achieve agreement. However, this agreement is not on the optimal solution. This can be shown by defining an accumulation state, $\widehat{\mathbf{x}}(k) = \sum_{i=1}^{n} [\boldsymbol{\pi}_r]_i \mathbf{x}_i(k)$, where $\boldsymbol{\pi}_r$ is the left eigenvector, corresponding the

eigenvalue of 1, of the row-stochastic matrix W, to obtain

$$\widehat{\mathbf{x}}(k+1) = \widehat{\mathbf{x}}(k) - \eta(k) \sum_{i=1}^{n} [\boldsymbol{\pi}_r]_i \nabla f_i\big(\mathbf{x}_i(k)\big). \tag{2.9}$$

It can be shown that the agents agree to the solution of the above iteration, which is suboptimal since the solution minimizes a weighted sum of the objective functions and not the sum. This argument leads to a modification of Eq. (2.9) that cancels the *imbalance* in the gradient caused by the fact that $\boldsymbol{\pi}_r$ is not a vector of all 1's, a consequence of losing the column-stochasticity in W. However, since we do not have prior knowledge on the left eigenvector of the row-stochastic weight matrix, we perform a push-sum (type) update to asymptotically learn its left eigenvector. The modification can be implemented as follows Mai and Abed, 2016:

$$\mathbf{x}_i(k+1) = \sum_{j=1}^{n} a_{ij}\mathbf{x}_j(k) - \eta(k) \frac{\nabla f_i\big(\mathbf{x}_i(k)\big)}{[\mathbf{y}_i(k)]_i}, \tag{2.10a}$$

$$\mathbf{y}_i(k+1) = \sum_{j=1}^{n} a_{ij}\mathbf{y}_j(k). \tag{2.10b}$$

The algorithm is initialized with $\mathbf{y}_i(0) = \mathbf{e}_i$ and arbitrary $\mathbf{x}_i(0)$'s. Eq. (2.10b) asymptotically learns the left eigenvector of the row-stochastic weight matrix $\underline{A} = \{a_{ij}\}$, i.e., $\lim_{k\to\infty} \mathbf{y}_i(k) = \boldsymbol{\pi}_r, \forall i$. The above algorithm achieves the convergence rate of $O(\frac{\ln k}{\sqrt{k}})$ under the Assumptions 1, 3, and 4, see Mai and Abed, 2016 for details.

Chapter 3

FROST (Fast Row-stochastic Optimization with uncoordinated STep-sizes)

3.1 Algorithm description

Based on the intuition in the previous chapter, we now describe FROST, i.e., a fast distributed algorithm based on row-stochastic weights and with non-identical step-sizes at the agents. Each agent i at the kth iteration maintains three state vectors, $\mathbf{x}_i(k)$, $\mathbf{y}_i(k)$ and $\mathbf{z}_i(k)$. At $k+1$-th iteration, the algorithm performs the following update:

$$\mathbf{x}_i(k+1) = \sum_{j=1}^{n} a_{ij}\mathbf{x}_j(k) - \eta_i \mathbf{z}_i(k), \tag{3.1a}$$

$$\mathbf{y}_i(k+1) = \sum_{j=1}^{n} a_{ij}\mathbf{y}_j(k), \tag{3.1b}$$

$$\mathbf{z}_i(k+1) = \sum_{j=1}^{n} a_{ij}\mathbf{z}_j(k) + \frac{\nabla f_i(\mathbf{x}_i(k+1))}{[\mathbf{y}_i(k+1)]_i} - \frac{\nabla f_i(\mathbf{x}_i(k))}{[\mathbf{y}_i(k)]_i}, \tag{3.1c}$$

where $\underline{A} = \{a_{ij}\}$ is a row-stochastic weight matrix and η_i's are uncoordinated step-sizes locally chosen at each agent. The algorithm is initialized with arbitrary $\mathbf{x}_i(0)$, $\mathbf{y}_i(0) = \mathbf{e}_i$, where $\mathbf{e}_i$ is a vector of 0's with a 1 at the ith location, and $\mathbf{z}_i(0) = \nabla f_i(\mathbf{x}_i(0))$. We point out that the initial condition for Eq. (3.1b) and the divisions

in Eq. (3.1c) require each agent to know and have a unique identifier; clearly, Assumption 4 is relevant here[1]. Note that Eq. (3.1c) is a modified gradient-estimation step, where the divisions are used to eliminate the imbalance caused by losing the column-stochasticity of the underlying weight matrix.

For analysis purposes, we write Eqs. (3.1) in a compact matrix form: To this aim, we denote by $\mathbf{x}^* \in \mathbb{R}^p$, the optimal solution of Problem P1, and define $\mathbf{x}(k), \mathbf{z}(k), \nabla\mathbf{f}(k) \in \mathbb{R}^{np}$ and $Y(k) \in \mathbb{R}^{n\times n}$ as follows:

$$\begin{aligned}
\mathbf{x}(k) &= [\mathbf{x}_1(k)^\top, \cdots, \mathbf{x}_n(k)^\top]^\top,\\
\mathbf{z}(k) &= [\mathbf{z}_1(k)^\top, \cdots, \mathbf{z}_n(k)^\top]^\top,\\
\nabla\mathbf{f}(k) &= [\nabla f_1\big(\mathbf{x}_1(k)\big)^\top, \cdots, \nabla f_n\big(\mathbf{x}_1(k)\big)^\top]^\top,\\
Y(k) &= [\mathbf{y}_1(k), \cdots, \mathbf{y}_n(k)]^\top,\\
\widetilde{Y}(k) &= \operatorname{diag}\big(Y(k)\big) \otimes I_p,\\
A &= \underline{A} \otimes I_p,\\
\boldsymbol{\eta} &= [\eta_1, \cdots, \eta_n]^\top,\\
D &= \operatorname{diag}\{\boldsymbol{\eta}\} \otimes I_p.
\end{aligned}$$

Recall that $\underline{A}$ is row-stochastic and $a_{ij} > 0, \forall (i,j) \in \mathcal{E}$. Given that the graph is strongly-connected and $\underline{A}$ is non-negative with positive diagonals, it is straightforward to verify that $\underline{A}$ is primitive and $\widetilde{Y}(k)$ is invertible for any k. Based on the notation above, Eq. (3.1) can be written equivalently as follows:

$$\mathbf{x}(k+1) = A\mathbf{x}(k) - D\mathbf{z}(k), \tag{3.2a}$$

$$Y(k+1) = AY(k), \tag{3.2b}$$

$$\mathbf{z}(k+1) = A\mathbf{z}(k) + \widetilde{Y}(k+1)^{-1}\nabla\mathbf{f}(k+1) - \widetilde{Y}(k)^{-1}\nabla\mathbf{f}(k), \tag{3.2c}$$

where $Y(0) = I_n$, $\mathbf{z}(0) = \nabla\mathbf{f}(0)$, and $\mathbf{x}(0)$ is arbitrary. We emphasize that the implementation of the algorithm needs no knowledge of agent's out-degree anywhere in the network in contrast to the earlier related work in Nedić and Olshevsky, 2015;

[1] If the agents do not have such knowledge, this requirement can be met rather easily with the help of, e.g., finite-time task allocation algorithms Kushner and Yin, 2003; Safavi and Khan, 2014, where the task at each agent is to pick a unique number from the set $\{1, \ldots, n\}$.

Tsianos, Lawlor, and Rabbat, 2012b; Xi and Khan, 2017b; Xi, Wu, and Khan, 2017; Xi, Xin, and Khan, 2017b; Xi and Khan, 2017a.

3.2 Convergence Analysis

In this section, we present the convergence analysis of FROST, i.e., Eqs. (4.6). We first define a few additional constants as follows:

$$\begin{aligned} Y_\infty &= \lim_{k\to\infty} Y(k) = \lim_{k\to\infty} \underline{A}^k, \\ \nabla \mathbf{f}^* &= [\nabla f_1(\mathbf{x}^*)^\top, \cdots, \nabla f_n(\mathbf{x}^*)^\top]^\top. \end{aligned}$$

We denote a few more constants as follows.

$$\begin{aligned} \tau &= \left|\!\left|\!\left| A - I_{np} \right|\!\right|\!\right|_2, \\ \epsilon &= \left|\!\left|\!\left| I_{np} - Y_\infty \right|\!\right|\!\right|_2, \\ \overline{\eta} &= \max_i\{\eta_i\}, \\ y &= \sup_k \left|\!\left|\!\left| Y(k) \right|\!\right|\!\right|_2, \\ \widetilde{y} &= \sup_k \left|\!\left|\!\left| \widetilde{Y}(k)^{-1} \right|\!\right|\!\right|_2. \end{aligned}$$

Since $\underline{A}$ is primitive and Y_0 is an $n \times n$ identity matrix, we have that $\{Y(k)\}$ is convergent, and all of its diagonal elements are nonzero and bounded, for all k. Therefore, y and $\widetilde{y}$ are finite.

3.2.1 Auxiliary relations

We now provide some lemmas that are useful in the rest of the paper.

Lemma 1. *Let Assumption 1 hold and consider the weight matrice $A = \underline{A} \otimes I_p$. There exists a vector norm, $\|\cdot\|$, such that for all $\mathbf{a} \in \mathbb{R}^{np}$,*

$$\|A\mathbf{a} - A_\infty \mathbf{a}\| \leq \sigma \|\mathbf{a} - A_\infty \mathbf{a}\|, \tag{3.3}$$

where $0 < \sigma < 1$ is some positive constant.

Proof. Since $\underline{A}$ is irreducible, row-stochastic with positive diagonals, from Perron-Frobenius theorem we have that $\rho(\underline{A}) = 1$, every eigenvalue of $\underline{A}$ other than 1 is strictly less than $\rho(\underline{A})$, and $\boldsymbol{\pi}_r^\top$ is a strictly positive left eigenvector corresponding to the eigenvalue of 1 with $\boldsymbol{\pi}_r^\top \mathbf{1}_n = 1$; thus $\lim_{k\to\infty} \underline{A}^k = \mathbf{1}_n \boldsymbol{\pi}_r^\top$. We further have

$$A_\infty = \lim_{k\to\infty} A^k = \left(\lim_{k\to\infty} \underline{A}^k\right) \otimes I_p = \left(\mathbf{1}_n \boldsymbol{\pi}_r^\top\right) \otimes I_p.$$

It follows that:

$$\begin{aligned} AA_\infty &= (\underline{A} \otimes I_p)\Big((\mathbf{1}_n \boldsymbol{\pi}_r^\top) \otimes I_p\Big) = A_\infty; \\ A_\infty A_\infty &= \Big((\mathbf{1}_n \boldsymbol{\pi}_r^\top) \otimes I_p\Big)\Big((\mathbf{1}_n \boldsymbol{\pi}_r^\top) \otimes I_p\Big) = A_\infty. \end{aligned}$$

Thus $AA_\infty - A_\infty A_\infty$ is a zero matrix, which leads to the following relation:

$$A\mathbf{a} - A_\infty \mathbf{a} = (A - A_\infty)(\mathbf{a} - A_\infty \mathbf{a}).$$

Next we note that

$$\rho(A - A_\infty) = \rho\Big(\left(\underline{A} - \mathbf{1}_n \boldsymbol{\pi}_r^\top\right) \otimes I_p\Big) < 1;$$

thus there exists a matrix norm, $\|\cdot\|$, with $\|A - A_\infty\| < 1$ and a compatible vector norm, $\|\cdot\|$, see Horn and Johnson, 2013: Chapter 5 for details, i.e.,

$$\|A\mathbf{a} - A_\infty \mathbf{a}\| \leq |||\, A - A_\infty \,||| \, \|\mathbf{a} - A_\infty \mathbf{a}\|,$$

and the lemma follows with $\sigma = |||\, A - A_\infty \,|||$. □

The lemma above regarding the contraction in the consensus process under some arbitrary norm is crucial throughout the convergence analysis. As shown above, the existence of some norm in which the consensus process with row-stochastic matrix A is a contraction does not follow directly from the standard 2-norm argument for doubly-stochastic matrices and is a non-trivial contribution of this work. Such and following related arguments built on this notion of contraction under arbitrary norms were first introduced in Xi, Xin, and Khan, 2017b for column-stochastic

weights and in Xi et al., 2018 for row-stochastic weights; these arguments are harmonized later to hold simultaneously for both row- and column-stochastic weights in the next chapter.

The next lemma is a standard result in the consensus and Markov chain theory. It says that the information diffusion speed over the network, characterized by the evolution of the row-stochastic weight matrix $\underline{A}$, is linearly fast.

Lemma 2. *(Horn et al. Horn and Johnson, 1990) Consider $Y(k)$, generated from the row-stochastic matrix, $\underline{A}$, and its limit Y_∞. There exists $0 < \gamma_1 < 1$ and some constant l such that*

$$|||\, Y(k) - Y_\infty \,|||_2 \leq l\gamma_1^k, \qquad \forall k. \tag{3.4}$$

As a consequence of Lemma 2, we establish the linear convergence of the sequences $\{\widetilde{Y}(k)^{-1}\}$ and $\{\widetilde{Y}(k+1)^{-1} - \widetilde{Y}(k)^{-1}\}$.

Lemma 3. *The following inequalities hold for all $k \geq 1$.*

1. $\left|\left|\left| \widetilde{Y}(k)^{-1} - \widetilde{Y}_\infty^{-1} \right|\right|\right|_2 \leq \sqrt{n}l\widetilde{y}^2\gamma_1^k$
2. $\left|\left|\left| \widetilde{Y}(k+1)^{-1} - \widetilde{Y}(k)^{-1} \right|\right|\right|_2 \leq 2\sqrt{n}l\widetilde{y}^2\gamma_1^k$

Proof. The proof of (a) is as follows:

$$\begin{aligned}
\left|\left|\left| \widetilde{Y}(k)^{-1} - \widetilde{Y}_\infty^{-1} \right|\right|\right|_2 &\leq \left|\left|\left| \widetilde{Y}(k)^{-1} \right|\right|\right|_2 \left|\left|\left| \widetilde{Y}(k) - \widetilde{Y}_\infty \right|\right|\right|_2 \left|\left|\left| \widetilde{Y}_\infty^{-1} \right|\right|\right|_2, \\
&\leq \widetilde{y}^2 \left|\left|\left| \left(\mathrm{diag}\big(Y(k)\big) - \mathrm{diag}\big(Y_\infty\big)\right) \otimes I_p \right|\right|\right|_2 \\
&\leq \widetilde{y}^2 \left\| Y(k) - Y_\infty \right\|_F \\
&\leq \sqrt{n}l\widetilde{y}^2\gamma_1^k,
\end{aligned}$$

where the last inequality uses Lemma 2 and the fact that $\|\cdot\|_F \leq |||\cdot|||_2$. The result in (b) is straightforward by applying (a), i.e.,

$$\begin{aligned}
\left|\left|\left| \widetilde{Y}(k+1)^{-1} - \widetilde{Y}(k)^{-1} \right|\right|\right|_2 \leq& \left|\left|\left| \widetilde{Y}(k+1)^{-1} - \widetilde{Y}_\infty^{-1} \right|\right|\right|_2 + \left|\left|\left| \widetilde{Y}_\infty^{-1} - \widetilde{Y}(k)^{-1} \right|\right|\right|_2, \\
\leq& \sqrt{n}l\widetilde{y}^2\gamma_1^{k+1} + \sqrt{n}l\widetilde{y}^2\gamma_1^k,
\end{aligned}$$

which completes the proof. □

The next lemma presents the dynamics that govern the evolution of the weighted sum of $\mathbf{z}(k)$; recall that $\mathbf{z}(k)$, in Eq. (3.2c), estimates the average of local gradients, $\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i(k))$.

Lemma 4. *The following equation holds for all k:*

$$Y_\infty \mathbf{z}(k) = Y_\infty \widetilde{Y}(k)^{-1}\nabla \mathbf{f}(k). \tag{3.5}$$

Proof. Noting that $Y_\infty A = Y_\infty$, we obtain from Eq. (3.2c)

$$Y_\infty \mathbf{z}(k) = Y_\infty \mathbf{z}(k-1) + Y_\infty \widetilde{Y}(k)^{-1}\nabla \mathbf{f}(k) - Y_\infty \widetilde{Y}(k-1)^{-1}\nabla \mathbf{f}(k-1).$$

Do this iteratively, and we have that

$$Y_\infty \mathbf{z}(k) = Y_\infty \mathbf{z}(0) + Y_\infty \widetilde{Y}(k)^{-1}\nabla \mathbf{f}(k) - Y_\infty \widetilde{Y}(0)^{-1}\nabla \mathbf{f}(0).$$

Because of the initial condition $\mathbf{z}(0) = \nabla \mathbf{f}(0)$ and $\widetilde{Y}(0) = I_n$, we have that $Y_\infty \mathbf{z}(0) = Y_\infty \widetilde{Y}(0)^{-1}\nabla \mathbf{f}(0)$, which completes the proof. □

The next lemma, a standard result in convex optimization theory from Bubeck, 2014; Qu and Li, 2017b, states that the distance to the optimal minimizer shrinks by at least a fixed ratio if we perform a gradient descent step.

Lemma 5. *Suppose that $g : \mathbb{R}^p \to \mathbb{R}$ is strongly convex with Lipschitz-continuous gradient. Let α and β be its strong-convexity and Lipschitz-continuity constants respectively. For $\forall \mathbf{x} \in \mathbb{R}^p$ and $0 < \delta < \frac{2}{\beta}$, we have*

$$\left\|\mathbf{x} - \delta\nabla g(\mathbf{x}) - \mathbf{x}^*\right\|_2 \leq \tau \left\|\mathbf{x} - \mathbf{x}^*\right\|_2,$$

where $\tau = \max\left(\left|1-\alpha\delta\right|, \left|1-\beta\delta\right|\right)$.

With the help of previous lemmas, we are ready to derive a crucial contraction relationship for the proposed algorithm.

3.2.2 Contraction relationship

Our strategy for the convergence proof is to bound $\|\mathbf{x}(k+1) - Y_\infty \mathbf{x}(k+1)\|$, $\|Y_\infty \mathbf{x}(k+1) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2$, and $\|\mathbf{z}(k+1) - Y_\infty \mathbf{z}(k+1)\|$ as a linear function of their values in the last iteration and $\nabla \mathbf{f}(k)$; this approach extends the work in Qu and Li, 2017c on doubly-stochastic weights to row-stochastic weights. In particular, if we can show that the (norm of the) vector of these three quantities goes to zero, then we can establish that $\mathbf{x}(k)$ goes to $\mathbf{1}_n \otimes \mathbf{x}^*$. We will present this contraction relationship in the next lemmas. First, we derive a bound for $\|\mathbf{x}(k+1) - Y_\infty \mathbf{x}(k+1)\|$, which is essentially the consensus error of the network.

Lemma 6. *The following inequality holds, $\forall k$:*

$$\|\mathbf{x}(k+1) - Y_\infty \mathbf{x}(k+1)\| \leq \sigma \|\mathbf{x}(k) - Y_\infty \mathbf{x}(k)\| + \overline{\eta} d \epsilon \|\mathbf{z}(k)\|_2, \tag{3.6}$$

where d is the equivalence-norm constant such that $\|\cdot\| \leq d\|\cdot\|_2$ and η is the maximum step-size among the agents. Recall $\|\cdot\|$ to be the vector norm introduced in Lemma 13.

Proof. Using Eq. (4.6a) and Lemma. 13, we have

$$\begin{aligned}
&\|\mathbf{x}(k+1) - Y_\infty \mathbf{x}(k+1)\| \\
&= \left\| A\mathbf{x}(k) - D\mathbf{z}(k) - Y_\infty \Big(A\mathbf{x}(k) - D\mathbf{z}(k) \Big) \right\|, \\
&\leq \sigma \|\mathbf{x}(k) - Y_\infty \mathbf{x}(k)\| + \|D\mathbf{z}(k) - Y_\infty D\mathbf{z}(k)\|, \\
&\leq \sigma \|\mathbf{x}(k) - Y_\infty \mathbf{x}(k)\| + d\|I_{np} - Y_\infty\|_2 |||\, D\, |||_2 \|\mathbf{z}(k)\|_2 \\
&\leq \sigma \|\mathbf{x}(k) - Y_\infty \mathbf{x}(k)\| + \overline{\eta} d \epsilon \|\mathbf{z}(k)\|_2,
\end{aligned}$$

which completes the proof. □

Next, we derive a bound for $\|Y_\infty \mathbf{x}(k+1) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2$. It is the error between the accumulation state of the network, $Y_\infty \mathbf{x}(k+1)$, and the global minimizer.

Lemma 7. *If* $0 < n\boldsymbol{\pi}_r^\top \boldsymbol{\eta} < \frac{2}{\beta}$, *the following inequality holds,* $\forall k$*:*

$$\begin{aligned}
&\|Y_\infty \mathbf{x}(k+1) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2 \\
&\leq \overline{\eta} n\beta c \|\mathbf{x}(k) - Y_\infty \mathbf{x}(k)\| + \lambda \|Y_\infty \mathbf{x}(k) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2 \\
&\quad + \overline{\eta} y c \|\mathbf{z}(k) - Y_\infty \mathbf{z}(k)\| + \overline{\eta}\sqrt{n} l y \widetilde{y}^2 \gamma_1^k \|\nabla \mathbf{f}(k)\|_2 ,
\end{aligned} \tag{3.7}$$

where $\lambda = \max\left(\left|1 - n\boldsymbol{\pi}_r^\top \boldsymbol{\eta}\alpha\right|, \left|1 - n\boldsymbol{\pi}_r^\top \boldsymbol{\eta}\beta\right|\right)$ *and* c *is the equivalence-norm constant such that* $\|\cdot\|_2 \leq c\|\cdot\|$.

Proof. Recalling that $Y_\infty = (\mathbf{1}_n \otimes I_p)(\boldsymbol{\pi}_r^\top \otimes I_p)$ and $Y_\infty A = Y_\infty$, We have the following:

$$\begin{aligned}
&\|Y_\infty \mathbf{x}(k+1) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2 \\
&= \left\| Y_\infty \Big(A\mathbf{x}(k) - D\mathbf{z}(k) + (D - D) Y_\infty \mathbf{z}(k) \Big) - \mathbf{1}_n \otimes \mathbf{x}^* \right\|_2 , \\
&\leq \|Y_\infty \mathbf{x}(k) - Y_\infty D Y_\infty \mathbf{z}(k) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2 + \overline{\eta} y c \|\mathbf{z}(k) - Y_\infty \mathbf{z}(k)\|.
\end{aligned} \tag{3.8}$$

Since the last term in the inequality above matches the second last term in Eq. (3.7), we only need to handle the first term. We further note that:

$$Y_\infty D Y_\infty = \Big((\mathbf{1}_n \boldsymbol{\pi}_r^\top) \otimes I_p \Big) \Big(\mathrm{diag}\{\boldsymbol{\eta}\} \otimes I_p \Big) \Big((\mathbf{1}_n \boldsymbol{\pi}_r^\top) \otimes I_p \Big) = (\boldsymbol{\pi}_r^\top \boldsymbol{\eta}) Y_\infty .$$

Now, we derive a upper bound for the first term in Eq. (3.8),

$$\begin{aligned}
&\|Y_\infty \mathbf{x}(k) - Y_\infty D Y_\infty \mathbf{z}(k) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2 \\
&\leq \left\| (\mathbf{1}_n \otimes I_p) \Big((\boldsymbol{\pi}_r^\top \otimes I_p)\mathbf{x}(k) - \mathbf{x}^* - n(\boldsymbol{\pi}_r^\top \boldsymbol{\eta}) \nabla F\big((\boldsymbol{\pi}_r^\top \otimes I_p)\mathbf{x}(k)\big) \Big) \right\|_2 \\
&\quad + \left\| n(\boldsymbol{\pi}_r^\top \boldsymbol{\eta})(\mathbf{1}_n \otimes I_p) \nabla F\big((\boldsymbol{\pi}_r^\top \otimes I_p)\mathbf{x}(k)\big) - (\boldsymbol{\pi}_r^\top \boldsymbol{\eta}) Y_\infty \mathbf{z}(k) \right\|_2 , \\
&:= s_1 + s_2 .
\end{aligned} \tag{3.9}$$

If $0 < n\boldsymbol{\pi}_r^\top \boldsymbol{\eta} < \frac{2}{\beta}$, according to Lemma 1 and 5, $s_1 \leq \lambda \|Y_\infty \mathbf{x}(k) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2$ and

$$\begin{aligned} s_2 &= (\boldsymbol{\pi}_r^\top \boldsymbol{\eta}) \left\| n(\mathbf{1}_n \otimes I_p)\nabla F\big((\boldsymbol{\pi}_r^\top \otimes I_p)\mathbf{x}(k)\big) - Y_\infty \mathbf{z}(k)\right\|_2, \\ &\leq \overline{\eta} \left\| n(\mathbf{1}_n \otimes I_p)\nabla F\big((\boldsymbol{\pi}_r^\top \otimes I_p)\mathbf{x}(k)\big) - (\mathbf{1}_n \otimes I_p)(\mathbf{1}_n^\top \otimes I_p)\nabla \mathbf{f}(k)\right\|_2, \\ &\quad + \overline{\eta} \left\| (\mathbf{1}_n \otimes I_p)(\mathbf{1}_n^\top \otimes I_p)\nabla \mathbf{f}(k) - Y_\infty \mathbf{z}(k)\right\|_2 \\ &:= s_3 + s_4. \end{aligned} \tag{3.10}$$

From Assumption 2, since the gradient of objective functions is Lipschitz-continuous, we have

$$s_3 \leq \overline{\eta} n \beta c \|\mathbf{x}(k) - Y_\infty \mathbf{x}(k)\|. \tag{3.11}$$

Note that $Y_\infty \cdot \widetilde{Y}_\infty^{-1} = (\mathbf{1}_n \otimes I_p)(\boldsymbol{\pi}_r^\top \otimes I_p) \cdot \left(\mathrm{diag}\big(Y_\infty\big)^{-1} \otimes I_p\right) = (\mathbf{1}_n \otimes I_p)(\mathbf{1}_n^\top \otimes I_p)$. According to Lemma 4, $Y_\infty \mathbf{z}(k) = Y_\infty \widetilde{Y}(k)^{-1}\nabla \mathbf{f}(k)$, we have:

$$s_4 = \overline{\eta} \left\| Y_\infty \widetilde{Y}_\infty^{-1}\nabla \mathbf{f}(k) - Y_\infty \widetilde{Y}(k)^{-1}\nabla \mathbf{f}(k)\right\|_2 \leq \overline{\eta}\sqrt{n} l y \widetilde{y}^2 \gamma_1^k \|\nabla \mathbf{f}(k)\|_2, \tag{3.12}$$

where the inequality above uses Lemma 3. Combining Eqs. (3.8)-(3.12), we finish the proof. □

Next, we bound $\|\mathbf{z}(k+1) - Y_\infty \mathbf{z}(k+1)\|$, the consensus error of gradient estimates at each agent.

Lemma 8. *The following inequality holds, $\forall k$:*

$$\begin{aligned} &\|\mathbf{z}(k+1) - Y_\infty \mathbf{z}(k+1)\| \\ &\leq \epsilon \widetilde{y}\beta\tau c d \|\mathbf{x}(k) - Y_\infty \mathbf{x}(k)\| + \sigma \|\mathbf{z}(k) - Y_\infty \mathbf{z}(k)\| + \epsilon \widetilde{y}\beta\overline{\eta} d \|\mathbf{z}(k)\|_2 \\ &\quad + 2d\sqrt{n} l \epsilon \widetilde{y}^2 \gamma_1^k \|\nabla \mathbf{f}(k)\|_2. \end{aligned} \tag{3.13}$$

Proof. According to Eq. (3.2c) and Lemma 1, we have

$$\begin{aligned} &\|\mathbf{z}(k+1) - Y_\infty \mathbf{z}(k+1)\| \\ &\leq \sigma \|\mathbf{z}(k) - Y_\infty \mathbf{z}(k)\| \\ &\quad + \left\| \left(\widetilde{Y}(k+1)^{-1}\nabla \mathbf{f}(k+1) - \widetilde{Y}(k)^{-1}\nabla \mathbf{f}(k)\right) - \left(Y_\infty \mathbf{z}(k+1) - Y_\infty \mathbf{z}(k)\right)\right\|, \end{aligned} \tag{3.14}$$

after expanding the first $\mathbf{z}(k+1)$ and adding and subtracting $Y_\infty \mathbf{z}(k)$. Note that $Y_\infty \mathbf{z}(k) = Y_\infty \widetilde{Y}(k)^{-1} \nabla \mathbf{f}(k)$ from Lemma 4. Therefore,

$$\begin{aligned}
&\left\| \left(\widetilde{Y}(k+1)^{-1} \nabla \mathbf{f}(k+1) - \widetilde{Y}(k)^{-1} \nabla \mathbf{f}(k) \right) - \left(Y_\infty \mathbf{z}(k+1) - Y_\infty \mathbf{z}(k) \right) \right\|_2 \\
&= \left\| \left(I_{np} - Y_\infty \right) \left(\widetilde{Y}(k+1)^{-1} \nabla \mathbf{f}(k+1) - \widetilde{Y}(k)^{-1} \nabla \mathbf{f}(k) \right) \right\|_2, \\
&\leq \epsilon \left\| \widetilde{Y}(k+1)^{-1} \nabla \mathbf{f}(k+1) - \widetilde{Y}(k+1)^{-1} \nabla \mathbf{f}(k) \right\|_2 \\
&\quad + \epsilon \left\| \widetilde{Y}(k+1)^{-1} \nabla \mathbf{f}(k) - \widetilde{Y}(k)^{-1} \nabla \mathbf{f}(k) \right\|_2, \\
&\leq \epsilon \widetilde{y} \beta \left\| \mathbf{x}(k+1) - \mathbf{x}(k) \right\|_2 + 2\sqrt{n} l \epsilon \widetilde{y}^2 \gamma_1^k \left\| \nabla \mathbf{f}(k) \right\|_2, \qquad (3.15)
\end{aligned}$$

where in the last inequality we use the Lipschitz-continuity of the objective functions and Lemma 3. We now bound $\|\mathbf{x}(k+1) - \mathbf{x}(k)\|_2$.

$$\begin{aligned}
\left\| \mathbf{x}(k+1) - \mathbf{x}(k) \right\|_2 &\leq \left\| (A - I_{np}) \mathbf{x}(k) \right\|_2 + \left\| \left\| D \right\| \right\|_2 \left\| \mathbf{z}(k) \right\|_2, \\
&\leq \left\| (A - I_{np}) \left(\mathbf{x}(k) - Y_\infty \mathbf{x}(k) \right) \right\|_2 + \overline{\eta} \left\| \mathbf{z}(k) \right\|_2, \\
&\leq \tau \left\| \mathbf{x}(k) - Y_\infty \mathbf{x}(k) \right\|_2 + \overline{\eta} \left\| \mathbf{z}(k) \right\|_2, \qquad (3.16)
\end{aligned}$$

where in the second inequality we use the fact that $(A - I_{np}) Y_\infty$ is a zero matrix. Combining Eqs. (3.14)-(3.16), we obtain the desired result. □

The last step is to bound $\|\mathbf{z}(k)\|_2$ in terms of $\|\mathbf{x}(k) - Y_\infty \mathbf{x}(k)\|$, $\|Y_\infty \mathbf{x}(k) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2$, and $\|\mathbf{z}(k) - Y_\infty \mathbf{z}(k)\|$. Then we can replace $\|\mathbf{z}(k)\|_2$ in Lemma 6-8 by this bound and a linear matrix inequality is completed.

Lemma 9. *The following inequality holds, $\forall k$:*

$$\begin{aligned}
\|\mathbf{z}(k)\|_2 \leq\ & cn\beta \|\mathbf{x}(k) - Y_\infty \mathbf{x}(k)\| + n\beta \|Y_\infty \mathbf{x}(k) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2 \\
& + d \|\mathbf{z}(k) - Y_\infty \mathbf{z}(k)\| + \sqrt{n} l y \widetilde{y}^2 \gamma_1^k \|\nabla \mathbf{f}(k)\|_2. \qquad (3.17)
\end{aligned}$$

Proof. Recall that $Y_\infty \widetilde{Y}_\infty^{-1} = (\mathbf{1}_n \otimes I_p)(\mathbf{1}_n^\top \otimes I_p)$ and $Y_\infty \mathbf{z}(k) = Y_\infty \widetilde{Y}(k)^{-1} \nabla \mathbf{f}(k)$ from Lemma 4. We have the following:

$$\begin{aligned}
\|\mathbf{z}(k)\|_2 &\leq \|\mathbf{z}(k) - Y_\infty \mathbf{z}(k)\|_2 + \|Y_\infty \mathbf{z}(k)\|_2 \\
&\leq d\|\mathbf{z}(k) - Y_\infty \mathbf{z}(k)\| + \|Y_\infty \widetilde{Y}(k)^{-1} \nabla \mathbf{f}(k) - Y_\infty \widetilde{Y}_\infty^{-1} \nabla \mathbf{f}(k)\|_2 \\
&\quad + \|Y_\infty \widetilde{Y}_\infty^{-1} \nabla \mathbf{f}(k) - (\mathbf{1}_n \otimes I_p)(\mathbf{1}_n^\top \otimes I_p) \nabla \mathbf{f}(\mathbf{x}^*)\|_2, \\
&\leq d\|\mathbf{z}(k) - Y_\infty \mathbf{z}(k)\| + \sqrt{n} l y \widetilde{y}^2 \gamma_1^k \|\nabla \mathbf{f}(k)\|_2 \\
&\quad + n\beta \|\mathbf{x}(k) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2, \\
&\leq cn\beta \|\mathbf{x}(k) - Y_\infty \mathbf{x}(k)\| + n\beta \|Y_\infty \mathbf{x}(k) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2 \\
&\quad + d\|\mathbf{z}(k) - Y_\infty \mathbf{z}(k)\| + \sqrt{n} l y \widetilde{y}^2 \gamma_1^k \|\nabla \mathbf{f}(k)\|_2,
\end{aligned} \tag{3.18}$$

where in the second inequality we use the fact that $(\mathbf{1}_n^\top \otimes I_p) \nabla \mathbf{f}(\mathbf{x}^*) = 0$, which is the optimality condition for Problem P1. □

Before the main result, we present an additional lemma from nonnegative matrix theory, which helps us to establish the condition for convergence and the convergence speed from the contraction relationship.

Lemma 10. *(Theorem 8.1.29 in Horn and Johnson, 2013) Let $X \in \mathbb{R}^{n \times n}$ be a nonnegative matrix and $\mathbf{x} \in \mathbb{R}^n$ be a positive vector. If $X\mathbf{x} < \omega \mathbf{x}$, then $\rho(X) < \omega$.*

3.2.3 Main results

With the help of the auxiliary relationships developed in the previous subsection, we are now able to present the main result as follows in Theorems 1 and 2. Theorem 1 says that the relationships we derive in the previous subsection indeed provides a contraction when the largest step-size, $\overline{\eta}$, is sufficiently small. Theorem 2 then establishes the linear rate of convergence of FROST.

Theorem 1. *Let Assumptions 1, 2 and 4 hold. If $0 < n\boldsymbol{\pi}_r^\top \boldsymbol{\eta} < \frac{2}{\beta}$, we have the following linear matrix inequality:*

$$\mathbf{t}(k+1) \leq J(\boldsymbol{\eta}) \mathbf{t}(k) + H(k) s(k), \ \forall k, \tag{3.19}$$

where $\mathbf{t}(k), \mathbf{s}(k) \in \mathbb{R}^3$ *and* $J(\boldsymbol{\eta}), H(k) \in \mathbb{R}^{3\times 3}$ *are defined as follows:*

$$\mathbf{t}(k) = \begin{bmatrix} \|\mathbf{x}(k) - Y_\infty \mathbf{x}(k)\| \\ \|Y_\infty \mathbf{x}(k) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2 \\ \|\mathbf{z}(k) - Y_\infty \mathbf{z}(k)\| \end{bmatrix}, \tag{3.20}$$

$$J(\boldsymbol{\eta}) = \begin{bmatrix} \sigma + a_1\overline{\eta} & a_2\overline{\eta} & a_3\overline{\eta} \\ a_4\overline{\eta} & \lambda & a_5\overline{\eta} \\ a_6 + a_7\overline{\eta} & a_8\overline{\eta} & \sigma + a_9\overline{\eta} \end{bmatrix}, \tag{3.21}$$

$$H(k) = \begin{bmatrix} \overline{\eta} d\epsilon\sqrt{n} l y \widetilde{y}^2 & 0 & 0 \\ \overline{\eta}\sqrt{n} l y \widetilde{y}^2 \gamma_1^k & 0 & 0 \\ d\sqrt{n} l \epsilon \widetilde{y}^2 \left(2 + \overline{\eta}\beta y \widetilde{y}\right) & 0 & 0 \end{bmatrix} \gamma_1^k, \tag{3.22}$$

$$\mathbf{s}(k) = \begin{bmatrix} \|\nabla \mathbf{f}(k)\|_2 \\ 0 \\ 0 \end{bmatrix}, \tag{3.23}$$

with the constants a_i*'s being*

$$\begin{aligned} a_1 &= cd\epsilon n\beta, & a_4 &= cn\beta & a_7 &= cdn\beta^2\epsilon\widetilde{y} \\ a_2 &= d\epsilon n\beta, & a_5 &= yc, & a_8 &= dn\beta^2\epsilon\widetilde{y} \\ a_3 &= d^2\epsilon, & a_6 &= \epsilon\widetilde{y}\beta\tau cd, & a_9 &= d^2\epsilon\beta\widetilde{y}, \end{aligned}$$

When the largest step-size, $\overline{\eta}$*, satisfies*

$$\overline{\eta} < \min\left\{ \frac{\delta_1(1-\sigma_A)}{a_1\delta_1 + a_2\delta_2 + a_3\delta_3}, \frac{(1-\sigma)\delta_3 - \delta_1 a_6}{a_7\delta_1 + a_8\delta_2 + a_9\delta_3}, \frac{1}{n\beta} \right\}, \tag{3.24}$$

where $\delta_1, \delta_2, \delta_3$ *are positive constants such that*

$$\delta_3 > 0, \qquad \delta_1 < \frac{(1-\sigma)\delta_3}{a_6}, \qquad \delta_2 > \frac{a_4\delta_1 + a_5\delta_3}{\alpha n [\boldsymbol{\pi}_r]_-}, \tag{3.25}$$

and $[\boldsymbol{\pi}_r]_-$ *is the smallest entry in* $\boldsymbol{\pi}_r$*, the spectral radius of* $J(\boldsymbol{\eta})$*,* $\rho(J(\boldsymbol{\eta}))$*, is strictly less than 1.*

Proof. Combining the results of Lemmas 6–9, one can verify that Eq. (3.19) holds if $0 < n\boldsymbol{\pi}_r^\top \boldsymbol{\eta} < \frac{2}{\beta}$. Recall that $\lambda = \max\left(\left|1 - \alpha n \boldsymbol{\pi}_r^\top \boldsymbol{\eta}\right|, \left|1 - \beta n \boldsymbol{\pi}_r^\top \boldsymbol{\eta}\right|\right)$. When $0 <$

$n\boldsymbol{\pi}_r^\top \boldsymbol{\eta} < \frac{1}{\beta}$, $\lambda = 1 - \alpha n \boldsymbol{\pi}_r^\top \boldsymbol{\eta}$, since $\alpha \leq \beta$; see, e.g., Bubeck, 2014 for details. In order to make $0 < n\boldsymbol{\pi}_r^\top \boldsymbol{\eta} < \frac{1}{\beta}$ hold, it is suffice to require $\overline{\eta} < \frac{1}{n\beta}$. The next step is to find an upper bound of the largest step-size, $\hat{\eta}$, such that $\rho(J(\boldsymbol{\eta})) < 1$ when $\overline{\eta} < \hat{\eta}$. In the light of Lemma 10, we solve for the range of the largest step-size, $\overline{\eta}$, and a positive vector $\boldsymbol{\delta} = [\delta_1, \delta_2, \delta_3]^\top$ from the following linear matrix inequality (entry-wise):

$$\begin{bmatrix} \sigma + a_1\overline{\eta} & a_2\overline{\eta} & a_3\overline{\eta} \\ a_4\overline{\eta} & 1 - \alpha n(\boldsymbol{\pi}_r^\top \boldsymbol{\eta}) & a_5\overline{\eta} \\ a_6 + a_7\overline{\eta} & a_8\overline{\eta} & \sigma + a_9\overline{\eta} \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix} < \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix}, \tag{3.26}$$

which is equivalent to the following set of inequalities:

$$\begin{cases} (a_1\delta_1 + a_2\delta_2 + a_3\delta_3)\overline{\eta} & < \delta_1(1-\sigma), \\ (a_4\delta_1 + a_5\delta_3)\overline{\eta} - \alpha n \boldsymbol{\pi}_r^\top \boldsymbol{\eta} & < 0, \\ (a_7\delta_1 + a_8\delta_2 + a_9\delta_3)\overline{\eta} & < (1-\sigma)\delta_3 - \delta_1 a_6. \end{cases} \tag{3.27}$$

Since the rightside of the third inequality in Eqs. (3.27) has to be positive, we have that:

$$0 < \delta_1 < \frac{(1-\sigma)\delta_3}{a_6} \tag{3.28}$$

In order to find the range of δ_2 such that the second inequality holds, it is suffice to solve the range of δ_2 such that the following inequality holds:

$$(a_4\delta_1 + a_5\delta_3)\overline{\eta} - \delta_2 \alpha n [\boldsymbol{\pi}_r]_- \overline{\eta} < 0,$$

where $[\boldsymbol{\pi}_r]_-$ is the smallest entry in $\boldsymbol{\pi}_r$. Therefore, as long as

$$\delta_2 > \frac{a_4\epsilon_1 + a_5\epsilon_3}{\alpha n [\boldsymbol{\pi}_r]_-}, \tag{3.29}$$

the second inequality in Eqs. (3.27) holds. The next step is to solve the range of $\overline{\eta}$ from the first and third inequality in Eqs. (3.27). We get

$$\overline{\eta} < \min \left\{ \frac{\delta_1(1-\sigma_A)}{a_1\delta_1 + a_2\delta_2 + a_3\delta_3}, \frac{(1-\sigma)\delta_3 - \delta_1 a_6}{a_7\delta_1 + a_8\delta_2 + a_9\delta_3} \right\},$$

where the range of δ_1 and δ_2 is given in Eq. (3.28) and Eq. (3.29) respectively and δ_3 is an arbitrary positive constant. The Theorem follows. □

We next prove the linear convergence of FROST to the global minimizer with the help of a few more auxiliary relations as follows.

Lemma 11. *Assume that the largest step-size among the agents satisfies the condition in Theorem 1. Then the following statements hold for all k,*

1. *for $0 < \gamma_1 < 1$, defined in Eq. (3.4), there exists a positive constant u_1 such that*

$$\left|\left|\left| H(k) \right|\right|\right|_2 = u_1 \gamma_1^k;$$

2. *there exists $0 < \gamma_2 < 1$ and a positive constant u_2, such that*

$$\left|\left|\left| J(\boldsymbol{\eta})^k \right|\right|\right|_2 \leq u_2 \gamma_2^k;$$

3. *Let $\gamma = \max\{\gamma_1, \gamma_2\}$ and $u = \dfrac{u_1 u_2}{\gamma}$. For all $0 \leq r \leq k-1$,*

$$\left|\left|\left| J(\boldsymbol{\eta})^{k-r-1} H(r) \right|\right|\right|_2 \leq u \gamma^k.$$

Proof. (a) Once can verify (a) according to Eq. (3.23), by letting

$$u_1 = \sqrt{(\overline{\eta} d \epsilon \sqrt{n} l y \widetilde{y}^2)^2 + (\overline{\eta} \sqrt{n} l y \widetilde{y}^2)^2 + (d \sqrt{n} l \epsilon \widetilde{y}^2)^2 \Big(2 + \overline{\eta} \beta y \widetilde{y}\Big)^2}.$$

(b) Since the spectral radius of $J(\boldsymbol{\eta})$ is strictly less than one, there exists some matrix norm of $J(\boldsymbol{\eta})$ also strictly less than one. We let the value of this matrix norm of $J(\boldsymbol{\eta})$ to be γ_2. Then, from the equivalence of norms, there exists a constant u_2 such that

$$\left|\left|\left| J(\boldsymbol{\eta})^k \right|\right|\right|_2 \leq u_2 \gamma_2^k. \tag{3.30}$$

Note that γ_2 can be infinitely close to the spectral radius of $J(\boldsymbol{\eta})$. See, e.g.,Horn and Johnson, 1990 for details.

(c) The proof of (c) follows from combining (a) and (b). □

The next Lemma is a standard result in convex optimization theory, which is useful in establishing convergence in many optimization algorithms.

Lemma 12. *(Polyak Polyak, 1987) If nonnegative sequences* $\{v(k)\}$, $\{q(k)\}$, $\{b(k)\}$ *and* $\{c(k)\}$ *are such that* $\sum_{k=0}^{\infty} b(k) < \infty$, $\sum_{k=0}^{\infty} c(k) < \infty$ *and*

$$v(k+1) \leq (1+b(k))v(k) - q(k) + c(k), \quad \forall t \geq 0,$$

then $\{v(k)\}$ *converges and* $\sum_{k=0}^{\infty} q(k) < \infty$.

To recap, we provide the linear iterative relation on $\mathbf{t}(k)$ with $J(\boldsymbol{\eta})$, $H(k)$ and $\mathbf{s}(k)$ in Theorem 1. We show that $\rho(J(\boldsymbol{\eta})) < 1$ and also show that $H(k)$ decays linearly in Theorem 1 and Lemma 11, respectively, if the largest step-size is sufficiently small. With the help of Lemma 12, we now present the linear convergence of FROST algorithm in the following theorem.

Theorem 2. *If the largest step-size,* $\overline{\eta}$, *satisfies the condition in Theorem 1, the sequence,* $\{\mathbf{x}(k)\}$, *generated by Eqs. (3.1), converges linearly to the optimal solution,* $\mathbf{1}_n \otimes \mathbf{x}^*$, *i.e., there exists a positive constant m such that*

$$\|\mathbf{x}(k) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2 \leq m(\gamma + \xi)^k, \qquad \forall k, \tag{3.31}$$

where ξ *is a arbitrarily small constant and* $\gamma = \max\{\gamma_1, \gamma_2\}$.

Proof. We iterate Eq. (3.19):

$$\mathbf{t}(k) \leq J(\boldsymbol{\eta})^k \mathbf{t}(0) + \sum_{r=0}^{k-1} J(\boldsymbol{\eta})^{k-r-1} H(r)\mathbf{s}(r). \tag{3.32}$$

Taking two-norm on both sides of the equation above, together with Lemma 11, we obtain that

$$\begin{aligned} \|\mathbf{t}(k)\|_2 &\leq \left\| J(\boldsymbol{\eta})^k \right\|_2 \|\mathbf{t}(0)\|_2 + \sum_{r=0}^{k-1} \left\| J(\boldsymbol{\eta})^{k-r-1} H(r) \right\|_2 \|\mathbf{s}(r)\|_2, \\ &\leq u_2 \gamma_2^k \|\mathbf{t}(0)\|_2 + \sum_{r=0}^{k-1} u\gamma^k \|\mathbf{s}(r)\|_2, \end{aligned} \tag{3.33}$$

in which we bound $\|\mathbf{s}(r)\|_2$ as

$$\begin{aligned}\|\mathbf{s}(r)\|_2 \leq & \|\nabla\mathbf{f}(r) - \nabla\mathbf{f}^*\|_2 + \|\nabla\mathbf{f}^*\|_2\,,\\ &\leq \beta\|\mathbf{x}(r) - Y_\infty\mathbf{x}(r)\|_2 + \beta\|Y_\infty\mathbf{x}(r) - \mathbf{1}_n\otimes\mathbf{x}^*\|_2 + \|\nabla\mathbf{f}^*\|_2\,,\\ &\leq (c+1)\beta\|\mathbf{t}(r)\|_2 + \|\nabla\mathbf{f}^*\|_2\,. \end{aligned} \tag{3.34}$$

Therefore, we have that for all k

$$\|\mathbf{t}(k)\|_2 \leq \Big(u_2\|\mathbf{t}(0)\|_2 + (c+1)u\beta\sum_{r=0}^{k-1}\|\mathbf{t}(r)\|_2 + uk\|\nabla\mathbf{f}^*\|_2\Big)\gamma^k. \tag{3.35}$$

Denote $v(k) = \sum_{r=0}^{k-1}\|\mathbf{t}(r)\|_2$, $h(k) = u_2\|\mathbf{t}(0)\|_2 + uk\|\nabla\mathbf{f}^*\|_2$, and $b = (c+1)u\beta$, then Eq. (3.35) can be written as

$$\|\mathbf{t}(k)\|_2 = v(k+1) - v(k) \leq \Big(h(k) + bv(k)\Big)\gamma^k, \tag{3.36}$$

which implies that $v(k+1) \leq (1+b\gamma^k)v(k) + h(k)\gamma^k$. Applying Lemma 12 with $b(k) = b\gamma^k$ and $c(k) = h(k)\gamma^k$ (here $q(k) = 0$), we have that $v(k)$ converges and thus bounded[2]. By Eq. (3.36), $\forall\mu \in (\gamma, 1)$ we have

$$\lim_{k\to\infty}\frac{\|\mathbf{t}(k)\|_2}{\mu^k} \leq \lim_{k\to\infty}\frac{\Big(h(k) + bv(k)\Big)\gamma^k}{\mu^k} = 0. \tag{3.37}$$

Therefore, $\|\mathbf{t}(k)\|_2 = O(\mu^k)$. In other words, there exists some positive constant ϕ such that for all k, we have:

$$\|\mathbf{t}(k)\|_2 \leq \phi(\gamma+\xi)^k, \tag{3.38}$$

where ξ is a arbitrarily small constant. It follows that

$$\begin{aligned}\|\mathbf{x}(k) - \mathbf{1}_n\otimes\mathbf{x}^*\|_2 \leq & \|\mathbf{x}(k) - Y_\infty\mathbf{x}(k)\|_2 + \|Y_\infty\mathbf{x}(k) - \mathbf{1}_n\otimes\mathbf{x}^*\|_2\,,\\ &\leq (c+1)\|\mathbf{t}(k)\|_2\,,\\ &\leq (c+1)\phi(\gamma+\xi)^k,\end{aligned}$$

which completes the prood. □

[2]In order to apply Lemma 12, we need to show that $\sum_{k=0}^{\infty}h(k)\gamma^k < \infty$, which follows from the fact that $\lim_{k\to\infty}\frac{h(k+1)\gamma^{k+1}}{h(k)\gamma^k} = \gamma < 1$.

3.3 Numerical Results

In this section, we use numerical experiments to verify our theoretical findings. We compare the performance of the proposed algorithm, FROST, with ADD-OPT Xi, Xin, and Khan, 2017b, see Section 2.2.2, and with the completely linear algorithm in Xin and Khan, Mar. 4th, 2018, which will be described in the next chapter. We adopt a simple uniform weighting strategy to construct the row- and column-stochastic weights when needed: $a_{ij} = 1/|\mathcal{N}_i^{\text{in}}|$, $b_{ij} = 1/|\mathcal{N}_j^{\text{out}}|$, $\forall i, j$.

3.3.1 Distributed linear estimation

A network of agents communicating over a directed graph seeks to estimate some input signal, $\mathbf{x}^* \in \mathbb{R}^p$. Each agent i measures an output signal through a linear function,

$$y_i = \mathbf{c}_i^\top \mathbf{x}^* + e_i,$$

where $\mathbf{c}_i \in \mathbb{R}^p, y_i \in \mathbb{R}, e_i \in \mathbb{R}$. The variables y_i and $\mathbf{c}_i$ are respectively the output signal and sensing matrix of agent i whereas e_i is some additive noise. Since the sensing matrix is rank 1 at each agent, no agent can recover $\mathbf{x}^*$ on its own and therefore cooperation among agents is essential. To obtain the optimal estimate of the input signal $\mathbf{x}^*$, the network of agents cooperatively solves the following convex optimization problem:

$$\min F(\mathbf{x}) = \sum_{i=1}^{n} \left(\mathbf{c}_i^\top \mathbf{x} - y_i \right)^2.$$

We set $n = 8$ and $p = 5$. The input signal and additive noise are randomly gen-

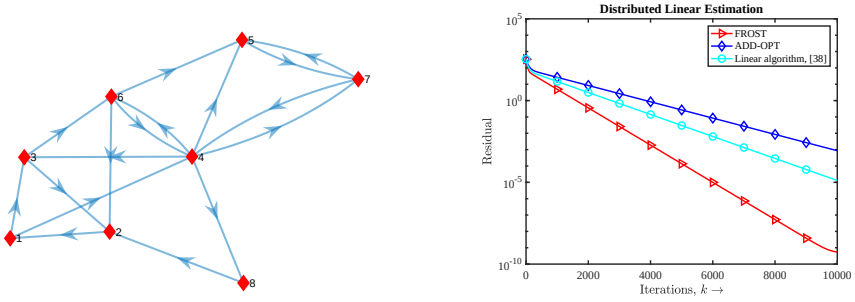

FIGURE 3.1: (Left) Strongly-connected and unbalanced directed graphs. (Right) Convergence comparison.

erated from standard uniform and Gaussian distribution respectively. The sensing matrices are generated from Gaussian distribution with zero mean and variance 5. We consider the network topology as the digraph shown in Fig. 3.1 (left) and the linear decaying of residuals are shown in Fig. 3.1 (right).

3.3.2 Distributed binary classification

Next, we consider a distributed binary classification problem, where we use logistic loss function to train a linear classifier. Each agent i has access to m_i training data, $(\mathbf{c}_{ij}, y_{ij}) \in \mathbb{R}^p \times \{-1, +1\}$, where $\mathbf{c}_{ij}$ contains p features of the jth training data at agent i and y_{ij} is the corresponding binary label. For privacy issues, agents do not share training data with each other. Therefore, in order to train a linear classifier from the entire data set, the network of agents cooperatively solves the following distributed logistic regression problem:

$$\min_{\mathbf{w}\in\mathbb{R}^p, b\in\mathbb{R}} F(\mathbf{w}, b) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} \ln\left[1 + \exp\left(-\left(\mathbf{w}^\top \mathbf{c}_{ij} + b\right) y_{ij}\right)\right] + \frac{n\lambda}{2}\|\mathbf{w}\|_2^2,$$

with each private loss function being

$$f_i(\mathbf{w}, b) = \sum_{j=1}^{m_i} \ln\left[1 + \exp\left(-\left(\mathbf{w}^\top \mathbf{c}_{ij} + b\right) y_{ij}\right)\right] + \frac{\lambda}{2}\|\mathbf{w}\|_2^2, \tag{3.39}$$

where $\frac{\lambda}{2}\|\mathbf{w}\|_2^2$ is a regularization term used to prevent overfitting of the data. The feature

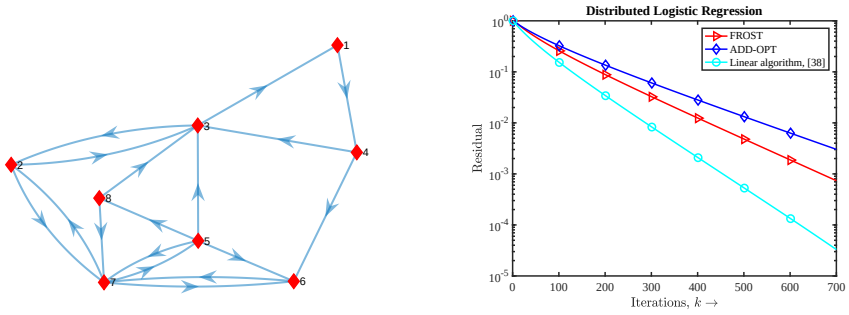

FIGURE 3.2: (Left) Strongly-connected and unbalanced directed graphs. (Right) Convergence comparison.

vectors, $\mathbf{c}_{ij}$'s, are randomly generated from Gaussian distribution with zero mean and variance 5. The binary labels are randomly generated from standard Bernoulli distribution. We consider the network topology as the digraph shown in Fig. 3.2 (left) and the linear decaying of residuals are shown in Fig. 3.2 (right).

Chapter 4

A linear algorithm with linear convergence

4.1 Algorithm description

In this chapter, we propose a completely linear algorithm with linear convergence over directed graphs. Each agent, $j \in \mathcal{V}$, maintains two variables: $\mathbf{x}_j(k)$, $\mathbf{y}_j(k) \in \mathbb{R}^p$, where k is discrete-time index. The algorithm, initialized with $\mathbf{y}_i(0) = \nabla f_i(\mathbf{x}_i(0)), \forall i$, and with arbitrary $\mathbf{x}_i(0)$, performs the following iterations.

$$\mathbf{x}_i(k+1) = \sum_{j \in \mathcal{N}_i^{\text{in}}} a_{ij}\mathbf{x}_j(k) - \eta \mathbf{y}_i(k), \tag{4.1a}$$

$$\mathbf{y}_i(k+1) = \sum_{j \in \mathcal{N}_i^{\text{in}}} b_{ij}\Big(\mathbf{y}_j(k) + \nabla f_j\big(\mathbf{x}_j(k+1)\big) - \nabla f_j\big(\mathbf{x}_j(k)\big)\Big), \tag{4.1b}$$

The weights, a_{ij}'s and b_{ij}'s satisfy the following conditions:

$$a_{ij} = \begin{cases} > 0, & j \in \mathcal{N}_i^{\text{in}}, \\ 0, & \text{otherwise}, \end{cases} \qquad \sum_{j=1}^{n} a_{ij} = 1, \forall i, \tag{4.2}$$

$$b_{ij} = \begin{cases} > 0, & i \in \mathcal{N}_j^{\text{out}}, \\ 0, & \text{otherwise}, \end{cases} \qquad \sum_{i=1}^{n} b_{ij} = 1, \forall j. \tag{4.3}$$

Eq. (4.2) leads to a row-stochastic matrix $\underline{A} = \{a_{ij}\}$, which is easy to implement as each agent locally decides the weights. Eq. (4.3), on the other hand, results in a column-stochastic matrix $\underline{B} = \{b_{ij}\}$, whose distributed implementation only requires each agent to know its out-degree. In particular, we can construct such weights as $b_{ij} = 1/|\mathcal{N}_j^{\text{out}}|, \forall i, j$.

The algorithm in Eqs. (4.1) can be explained as follows. To implement Eq. (4.1a), the receiving agent i decides on the weights a_{ij} assigned to the incoming $\mathbf{x}_j(k)$'s such that a_{ij}'s sum to 1. The step-size, η, is some positive constant. Implementation of Eq. (4.1b) requires the sending agent to scale the transmission $\mathbf{y}_j(k) + \mathbf{r}_j(k)$ by appropriate choice of b_{ij}'s (to ensure column-stochasticity of $\underline{B}$) as the out-degree of agent j may not be known to agent i. Agent i subsequently adds these received messages to implement Eq. (4.1b). Intuitively, Eq. (4.1b) asymptotically learns the average, $\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i(k))$, of the local gradients, Xu et al., 2015; Qu and Li, 2017b; Qu and Li, 2017a; Zhu and Martínez, 2010; and thus Eq. (4.1a) approaches a centralized gradient descent, as the descent direction, $\mathbf{y}_i(k)$, becomes the gradient of the global objective function over time.

4.2 Relation with existing work

We now briefly compare the proposed algorithm with existing techniques. As introduced in the previous chapter, the algorithms in Refs. Qu and Li, 2017b; Xu et al., 2015, can be summarized as a single class of algorithms over undirected graphs with the following form:

$$\mathbf{x}_i(k+1) = \sum_{j=1}^{n} w_{ij}\mathbf{x}_j(k) - \eta\mathbf{y}_i(k), \tag{4.4a}$$

$$\mathbf{y}_i(k+1) = \sum_{j=1}^{n} w_{ij}\mathbf{y}_j(k) + \nabla f_i\big(\mathbf{x}_i(k+1)\big) - \nabla f_i\big(\mathbf{x}_i(k)\big), \tag{4.4b}$$

where $W = \{w_{ij}\}$ is doubly-stochastic. This algorithm, however, is not applicable to directed graphs since it may not be possible to construct doubly-stochastic weights.

To overcome this issue, ref. Xi, Xin, and Khan, 2017a; Nedić, Olshevsky, and Shi, 2017 propose the following algorithm:

$$\begin{aligned}
\mathbf{x}_i(k+1) &= \sum_{j=1}^{n} b_{ij}\mathbf{x}_j(k) - \eta\mathbf{w}_i(k),\\
y_i(k+1) &= \sum_{j=1}^{n} b_{ij}y_j(k),\\
\mathbf{z}_i(k+1) &= \frac{\mathbf{x}_i(k+1))}{y_i(k+1)}.\\
\mathbf{w}_i(k+1) &= \sum_{j=1}^{n} b_{ij}\mathbf{w}_j(k) + \nabla f_i\big(\mathbf{z}_i(k+1)\big) - \nabla f_i(\mathbf{z}_i(k)),
\end{aligned}$$

where $\underline{B} = \{b_{ij}\}$ is column-stochastic.

Another method, refs. Xi et al., 2018; Xin, Xi, and Khan, 2018 proposes the following algorithm:

$$\begin{aligned}
\mathbf{y}_i(k+1) &= \sum_{j=1}^{n} a_{ij}\mathbf{y}_i(k), \\
\mathbf{x}_i(k+1) &= \sum_{j=1}^{n} a_{ij}\mathbf{x}_i(k) - \eta_i\mathbf{z}_i(k), \\
\mathbf{z}_i(k+1) &= \sum_{j=1}^{n} a_{ij}\mathbf{z}_i(k) + \frac{\nabla f_i\big(\mathbf{x}_i(k+1)\big)}{[\mathbf{y}_i(k+1)]_i} - \frac{\nabla f_i\big(\mathbf{x}_i(k)\big)}{[\mathbf{y}_i(k)]_i},
\end{aligned}$$

where $\underline{A} = \{a_{ij}\}$ is a row-stochastic.

Note that these algorithms leverage push-sum (type) techniques which use an independent algorithm to asymptotically learn either the left or right eigenvector, corresponding to the eigenvalue of 1, of the weight matrix. However, it adds nonlinearity to the overall algorithm along with additional computation and communication costs in contrast to the proposed algorithm in Eqs. (4.1).

Remarks: The algorithm, Eqs. (4.1), proposed in this thesis can be viewed as related to Eq. (4.4) but without doubly-stochastic weights, due to which we lose the nice eigenstructure within the weight matrices. It is rather straightforward to notice that a linear extension of Eqs. (4.4) to the directed graphs is non-trivial as all earlier attempts were made by adding nonlinearity to the original set of equations. One of the major challenges lies in the fact that even though the contraction of a doubly-stochastic W is well-established in the subspace orthogonal to $\mathbf{1}_n$, it is not straightforward to establish simultaneous contractions for a row-stochastic matrix, $\underline{A}$, and a column-stochastic matrix, $\underline{B}$. The latter requires working with arbitrary norms (as opposed to the 2-norm applicable to doubly-stochastic matrices) and norm-equivalence constants, as we show in Lemma 13 and onwards.

4.3 Convergence Analysis

For the sake of analysis, we now write Eqs. (4.1) in matrix form. The variables $\mathbf{x}(k)$ and $\mathbf{y}(k)$ collect all the local variables $\mathbf{x}_i(k)$'s and $\mathbf{y}_i(k)$'s in a vector, respectively, and

$$\nabla\mathbf{f}(k) = \begin{bmatrix} \nabla f_1\big(\mathbf{x}_1(k)\big) \\ \vdots \\ \nabla f_n\big(\mathbf{x}_n(k)\big) \end{bmatrix} \in \mathbb{R}^{np}. \tag{4.5}$$

Let $A = \underline{A} \otimes I_p$ and $B = \underline{B} \otimes I_p$, where $\otimes$ is the Kronecker product. We denote $\mathbf{x}^*$ as the optimal solution of Problem P1. We now rewrite Eqs. (4.1) in a compact matrix form as

follows:

$$\mathbf{x}(k+1) = A\mathbf{x}(k) - \eta\mathbf{y}(k), \tag{4.6a}$$

$$\mathbf{y}(k+1) = B\Big(\mathbf{y}(k) + \nabla\mathbf{f}(k+1) - \nabla\mathbf{f}(k)\Big), \tag{4.6b}$$

where $\mathbf{y}(0) = \nabla\mathbf{f}(0)$ and $\mathbf{x}(0)$ is arbitrary.

4.3.1 Auxiliary relations

We next start the convergence analysis with a key lemma regarding the contraction in consensus process with row- and column-stochastic weight matrices, respectively.

Lemma 13. *Let Assumption 1 hold and consider the weight matrices $A = \underline{A} \otimes I_p$ and $B = \underline{B} \otimes I_p$. Then there exist vector norms, $\|\cdot\|_A$ and $\|\cdot\|_B$, such that for all $\mathbf{a} \in \mathbb{R}^{np}$,*

$$\|A\mathbf{a} - A_\infty\mathbf{a}\|_A \leq \sigma_A \|\mathbf{a} - A_\infty\mathbf{a}\|_A, \tag{4.7}$$

$$\|B\mathbf{a} - B_\infty\mathbf{a}\|_B \leq \sigma_B \|\mathbf{a} - B_\infty\mathbf{a}\|_B, \tag{4.8}$$

where $0 < \sigma_A < 1$ and $0 < \sigma_B < 1$ are some constants.

Proof. Since $\underline{A}$ is irreducible, row-stochastic with positive diagonals, from Perron-Frobenius theorem we have that $\rho(\underline{A}) = 1$, every eigenvalue of $\underline{A}$ other than 1 is strictly less than $\rho(\underline{A})$, and $\boldsymbol{\pi}_r^\top$ is a strictly positive left eigenvector corresponding to the eigenvalue of 1 with $\mathbf{1}_n^\top \boldsymbol{\pi}_r = 1$; thus $\lim_{k\to\infty} \underline{A}^k = \mathbf{1}_n\boldsymbol{\pi}_r^\top$. We further have

$$A_\infty = \lim_{k\to\infty} A^k = \left(\lim_{k\to\infty} \underline{A}^k\right) \otimes I_p = \left(\mathbf{1}_n\boldsymbol{\pi}_r^\top\right) \otimes I_p.$$

It follows that

$$AA_\infty = (\underline{A} \otimes I_p)\Big((\mathbf{1}_n\boldsymbol{\pi}_r^\top) \otimes I_p\Big) = A_\infty,$$

$$A_\infty A_\infty = \Big((\mathbf{1}_n\boldsymbol{\pi}_r^\top) \otimes I_p\Big)\Big((\mathbf{1}_n\boldsymbol{\pi}_r^\top) \otimes I_p\Big) = A_\infty.$$

Thus $AA_\infty - A_\infty A_\infty$ is a zero matrix, which leads to the following relation:

$$A\mathbf{a} - A_\infty\mathbf{a} = (A - A_\infty)(\mathbf{a} - A_\infty\mathbf{a}). \tag{4.9}$$

Since $\rho(A - A_\infty) = \rho((\underline{A} - \mathbf{1}_n \boldsymbol{\pi}_r^\top) \otimes I_p) < 1$, we have from Lemma 5.6.10 in Horn and Johnson, 2013 that there exists a matrix norm, say $|||\cdot|||_A$, such that

$$\sigma_A \triangleq |||\, A - A_\infty \,|||_A < 1. \tag{4.10}$$

Moreover, from Theorem 5.7.13 in Horn and Johnson, 2013, we know that for any matrix norm, $|||\cdot|||_A$, there exists a compatible vector norm, say $\|\cdot\|_A$, such that $\|X\mathbf{x}\|_A \leq |||\, X \,|||_A \|\mathbf{x}\|_A$, for all matrices, X, and all vectors, $\mathbf{x}$; hence, Eq. (4.9) leads to

$$\begin{aligned} \|A\mathbf{a} - A_\infty \mathbf{a}\|_A &= \|(A - A_\infty)(\mathbf{a} - A_\infty \mathbf{a})\|_A, \\ &\leq |||\, A - A_\infty \,|||_A \|\mathbf{a} - A_\infty \mathbf{a}\|_A, \\ &= \sigma_A \|\mathbf{a} - A_\infty \mathbf{a}\|_A, \end{aligned}$$

and Eq. (4.7) follows, while Eq. (4.8) follows similarly for some matrix norm, $|||\cdot|||_B$, with $\sigma_B \triangleq |||\, B - B_\infty \,|||_B$. □

The following lemma is a direct consequence of the column-stochasticity of $\underline{B}$ and the initial condition that $\mathbf{y}(0) = \nabla \mathbf{f}(0)$.

Lemma 14. *We have* $(\mathbf{1}_n^\top \otimes I_p)\mathbf{y}(k) = (\mathbf{1}_n^\top \otimes I_p)\nabla \mathbf{f}(k), \forall k$.

Proof. Recall Eq. (4.6b) and multiply both sides of Eq. (4.6b) with $\mathbf{1}_n^\top \otimes I_p$. We get

$$\begin{aligned} &(\mathbf{1}_n^\top \otimes I_p)\mathbf{y}(k+1) \\ &= (\mathbf{1}_n^\top \otimes I_p)(\underline{B} \otimes I_p)\Big(\mathbf{y}(k) + \nabla \mathbf{f}(k+1) - \nabla \mathbf{f}(k)\Big) \\ &= (\mathbf{1}_n^\top \otimes I_p)\mathbf{y}(k) + (\mathbf{1}_n^\top \otimes I_p)\nabla \mathbf{f}(k+1) - (\mathbf{1}_n^\top \otimes I_p)\nabla \mathbf{f}(k) \\ &= (\mathbf{1}_n^\top \otimes I_p)\Big(\mathbf{y}(0) - \nabla \mathbf{f}(0)\Big) + (\mathbf{1}_n^\top \otimes I_p)\nabla \mathbf{f}(k+1) \\ &= (\mathbf{1}_n^\top \otimes I_p)\nabla \mathbf{f}(k+1), \end{aligned}$$

which completes the proof. □

Lemma 14 shows that the average of $\mathbf{y}_i(k)$'s preserves the average of local gradients. The subsequent convergence analysis is based on deriving a contraction relationship in the proposed algorithm, i.e., $\|\mathbf{x}(k+1) - A_\infty \mathbf{x}(k+1)\|_A$, $\|A_\infty \mathbf{x}(k+1) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2$, and $\|\mathbf{y}(k+1) - B_\infty \mathbf{y}(k+1)\|_B$, are bounded linearly by their values in the last iteration. We capture a relationship on these objects in the next lemmas. Before we proceed, note that all vector norms on finite-dimensional vector space are equivalent, i.e., there exist finite and positive

constants, c, d, h, l, g, m, such that:

$$\|\cdot\|_A \leq c\|\cdot\|_B, \quad \|\cdot\|_2 \leq h\|\cdot\|_B, \quad \|\cdot\|_2 \leq g\|\cdot\|_A,$$
$$\|\cdot\|_B \leq d\|\cdot\|_A, \quad \|\cdot\|_B \leq l\|\cdot\|_2, \quad \|\cdot\|_A \leq m\|\cdot\|_2.$$

Lemma 15. *The following inequality holds, $\forall k$:*

$$\begin{aligned}&\|\mathbf{x}(k+1) - A_\infty \mathbf{x}(k+1)\|_A \\ &\leq \sigma_A \|\mathbf{x}(k) - A_\infty \mathbf{x}(k)\|_A + \eta m \|\!| I_{np} - A_\infty |\!\|_2 \, \|\mathbf{y}(k)\|_2\end{aligned}$$

Proof. Using Eq. (4.6a) and Lemma 13, we have

$$\begin{aligned}&\|\mathbf{x}(k+1) - A_\infty \mathbf{x}(k+1)\|_A \\ &= \|A\mathbf{x}(k) - \eta \mathbf{y}(k) - A_\infty \Big(A\mathbf{x}(k) - \eta \mathbf{y}(k)\Big)\|_A, \\ &\leq \sigma_A \|\mathbf{x}(k) - A_\infty \mathbf{x}(k)\|_A + \eta m \|\mathbf{y}(k) - A_\infty \mathbf{y}(k)\|_2, \\ &\leq \sigma_A \|\mathbf{x}(k) - A_\infty \mathbf{x}(k)\|_A + \eta m \|\!| I_{np} - A_\infty |\!\|_2 \, \|\mathbf{y}(k)\|_2\end{aligned}$$

and the lemma follows. □

Next, we develop a relation for $\|A_\infty \mathbf{x}(k+1) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2$.

Lemma 16. *The following holds, $\forall k$, when $0 < \eta < \frac{2}{n\beta \boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c}$:*

$$\begin{aligned}&\|A_\infty \mathbf{x}(k+1) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2 \\ &\leq \eta n \beta g (\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c) \|\mathbf{x}(k) - A_\infty \mathbf{x}(k)\|_A \\ &\quad + \lambda \|A_\infty \mathbf{x}(k) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2 + \eta h \|\!| A_\infty |\!\|_2 \|\mathbf{y}(k) - B_\infty \mathbf{y}(k)\|_B, \end{aligned} \tag{4.11}$$

where $\lambda = \max\left(\left|1 - \alpha n \eta (\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c)\right|, \left|1 - \beta n \eta (\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c)\right|\right)$.

Proof. With $A_\infty = \mathbf{1}_n \boldsymbol{\pi}_r^\top \otimes I_p = (\mathbf{1}_n \otimes I_p)(\boldsymbol{\pi}_r^\top \otimes I_p)$,

$$A_\infty B_\infty = (\mathbf{1}_n \boldsymbol{\pi}_r^\top \otimes I_p)(\boldsymbol{\pi}_c \mathbf{1}_n^\top \otimes I_p) = \boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c (\mathbf{1}_n \mathbf{1}_n^\top \otimes I_p),$$

and recalling Eq. (4.6a), we have

$$\begin{aligned}
&\|A_\infty \mathbf{x}(k+1) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2 \\
&= \|A_\infty \Big(A\mathbf{x}(k) - \eta \mathbf{y}(k) + B_\infty \mathbf{y}(k)(-\eta + \eta) \Big) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2, \\
&\leq \|(\mathbf{1}_n \boldsymbol{\pi}_r^\top \otimes I_p)\mathbf{x}(k) - (\mathbf{1}_n \otimes I_p)\mathbf{x}^* - \eta A_\infty B_\infty \mathbf{y}(k)\|_2 \\
&\quad + \eta h |\!|\!| A_\infty |\!|\!|_2 \|\mathbf{y}(k) - B_\infty \mathbf{y}(k)\|_B. \qquad (4.12)
\end{aligned}$$

Since the last term above matches with the last term in Eq. (4.11), what is left is to manipulate the first term. Before we proceed, define $\nabla F(k) = \nabla F\big((\boldsymbol{\pi}_r^\top \otimes I_p)\mathbf{x}(k)\big)$, which is the global gradient evaluated at $(\boldsymbol{\pi}_r^\top \otimes I_p)\mathbf{x}(k)$. We have

$$\begin{aligned}
&\|(\mathbf{1}_n \boldsymbol{\pi}_r^\top \otimes I_p)\mathbf{x}(k) - (\mathbf{1}_n \otimes I_p)\mathbf{x}^* - \eta A_\infty B_\infty \mathbf{y}(k)\|_2 \\
&\leq \Big\| (\mathbf{1}_n \otimes I_p)\Big((\boldsymbol{\pi}_r^\top \otimes I_p)\mathbf{x}(k) - \mathbf{x}^* - n\eta (\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c)\nabla F(k) \Big) \Big\|_2 \\
&\quad + \eta (\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c) \Big\| n(\mathbf{1}_n \otimes I_p)\nabla F(k) - (\mathbf{1}_n, \otimes I_p)(\mathbf{1}_n^\top \otimes I_p)\mathbf{y}(k) \Big\|_2, \\
&:= s_1 + \eta s_2.
\end{aligned}$$

From Lemma 5, we have that if $0 < \eta < 2/(n\beta \boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c)$,

$$s_1 \leq \lambda \|A_\infty \mathbf{x}(k) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2.$$

Recall that $(\mathbf{1}_n^\top \otimes I_p)\mathbf{y}(k) = (\mathbf{1}_n^\top \otimes I_p)\nabla \mathbf{f}(k), \forall k$, from Lemma 14, we have

$$s_2 \leq n\beta g (\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c)\|\mathbf{x}(k) - A_\infty \mathbf{x}(k)\|_A.$$

The lemma follows by using the above bounds in Eq. (4.12). □

Next, we develop a relation for $\|\mathbf{y}(k+1) - B_\infty \mathbf{y}(k+1)\|_B$.

Lemma 17. *The following inequality holds, $\forall k$:*

$$\begin{aligned}
&\|\mathbf{y}(k+1) - B_\infty \mathbf{y}(k+1)\|_B \\
&\leq \sigma_B \beta l g |\!|\!| A - I_{np} |\!|\!|_2 \|\mathbf{x}(k) - A_\infty \mathbf{x}(k)\|_A \\
&\quad + \sigma_B \|\mathbf{y}(k) - B_\infty \mathbf{y}(k)\|_B + \eta \sigma_B \beta l \|\mathbf{y}(k)\|_2. \qquad (4.13)
\end{aligned}$$

Proof. We note that

$$\begin{aligned}
&\|\mathbf{y}(k+1) - B_\infty \mathbf{y}(k+1)\|_B \\
&= \left\| B\Big(\mathbf{y}(k) + \nabla \mathbf{f}(k+1) - \nabla \mathbf{f}(k)\Big) - B_\infty B\Big(\mathbf{y}(k) + \nabla \mathbf{f}(k+1) - \nabla \mathbf{f}(k)\Big) \right\|_B, \\
&\leq \sigma_B \|\mathbf{y}(k) - B_\infty \mathbf{y}(k)\|_B + \sigma_B \beta l \|\mathbf{x}(k+1) - \mathbf{x}(k)\|_2, \qquad (4.14)
\end{aligned}$$

because of Lemma 13. Now we analyze $\|\mathbf{x}(k+1) - \mathbf{x}(k)\|_2$.

$$\begin{aligned}
&\|\mathbf{x}(k+1) - \mathbf{x}(k)\|_2 \\
&= \|A\mathbf{x}(k) - \eta \mathbf{y}(k) - \mathbf{x}(k)\|_2, \\
&= \left\| (A - I_{np})\big(\mathbf{x}(k) - A_\infty \mathbf{x}(k)\big) - \eta \mathbf{y}(k) \right\|_2, \\
&\leq |||\, A - I_{np} \,|||_2 g \|\mathbf{x}(k) - A_\infty \mathbf{x}(k)\|_A + \eta \|\mathbf{y}(k)\|_2. \qquad (4.15)
\end{aligned}$$

The lemma follows by plugging Eq. (4.15) into Eq. (4.14). □

The last step is to bound $\|\mathbf{y}(k)\|_2$ in terms of $\|\mathbf{x}(k) - A_\infty \mathbf{x}(k)\|_A$, $\|A_\infty \mathbf{x}(k) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2$, and $\|\mathbf{y}(k) - B_\infty \mathbf{y}(k)\|_B$. Then we can replace $\|\mathbf{y}(k)\|_2$ in Lemma 15-17 by this bound to complete the contraction relationship.

Lemma 18. *The following inequality holds, $\forall k$:*

$$\begin{aligned}
\|\mathbf{y}(k)\|_2 \leq\ & g\beta |||\, B_\infty \,|||_2 \|\mathbf{x}(k) - A_\infty \mathbf{x}(k)\|_A \\
&+\ \beta |||\, B_\infty \,|||_2 \|A_\infty \mathbf{x}(k) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2 +\ h\|\mathbf{y}(k) - B_\infty \mathbf{y}(k)\|_B. \qquad (4.16)
\end{aligned}$$

Proof. Recall that $B_\infty = (\boldsymbol{\pi}_c \otimes I_p)(\mathbf{1}_n^\top \otimes I_p)$. We have

$$\|\mathbf{y}(k)\|_2 \leq h\|\mathbf{y}(k) - B_\infty \mathbf{y}(k)\|_B + \|B_\infty \mathbf{y}(k)\|_2. \qquad (4.17)$$

We next bound $\|B_\infty \mathbf{y}(k)\|_2$:

$$\begin{aligned}
\|B_\infty \mathbf{y}(k)\|_2 &= \|(\boldsymbol{\pi}_c \otimes I_p)(\mathbf{1}_n^\top \otimes I_p)\mathbf{y}(k)\|_2 \\
&= \|\boldsymbol{\pi}_c\|_2 \|(\mathbf{1}_n^\top \otimes I_p)\nabla \mathbf{f}(k)\|_2 \\
&= \|\boldsymbol{\pi}_c\|_2 \left\| \sum_{i=1}^n \nabla f_i(\mathbf{x}_i(k)) - \sum_{i=1}^n \nabla f_i(\mathbf{x}^*) \right\|_2 \\
&\leq \|\boldsymbol{\pi}_c\|_2 \beta \sum_{i=1}^n \|\mathbf{x}_i(k) - \mathbf{x}^*\|_2 \\
&\leq \|\boldsymbol{\pi}_c\|_2 \beta \sqrt{n} \|\mathbf{x}(k) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2, \\
&\leq |\!|\!| B_\infty |\!|\!|_2 \beta g \|\mathbf{x}(k) - A_\infty \mathbf{x}(k)\|_A + |\!|\!| B_\infty |\!|\!|_2 \beta \|A_\infty \mathbf{x}(k) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2, \qquad (4.18)
\end{aligned}$$

where the second last inequality uses Jensen's inequality and the last inequality uses the fact that $|\!|\!| B_\infty |\!|\!|_2 = \sqrt{n}\|\boldsymbol{\pi}_c\|_2$. The lemma follows by plugging Eqs. (4.18) into Eq. (4.17). □

4.3.2 Main results

With the help of auxiliary relations developed in the previous subsection, we now present the main result, which establishes the geometric convergence of the proposed algorithm.

Theorem 3. *Let Assumptions 1, 2 and 5 hold. If* $\eta < \frac{2}{n\beta \boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c}$, *we have the following linear matrix inequality (entry-wise):*

$$\mathbf{t}(k+1) \leq J(\eta)\mathbf{t}(k), \ \forall k, \qquad (4.19)$$

where $\mathbf{t}(k) \in \mathbb{R}^3$ *and* $J(\eta) \in \mathbb{R}^{3\times 3}$ *are defined as follows:*

$$\mathbf{t}(k) = \begin{bmatrix} \|\mathbf{x}(k) - A_\infty \mathbf{x}(k)\|_A \\ \|A_\infty \mathbf{x}(k) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2 \\ \|\mathbf{y}(k) - B_\infty \mathbf{y}(k)\|_B \end{bmatrix}, \qquad (4.20)$$

$$J(\eta) = \begin{bmatrix} \sigma_A + a_1\eta & a_2\eta & a_3\eta \\ a_4\eta & \lambda & a_5\eta \\ a_6 + a_7\eta & a_8\eta & \sigma_B + a_9\eta \end{bmatrix}, \qquad (4.21)$$

with the positive constants a_i's being

$$\begin{aligned}
a_1 &= mg\beta |||\, I_{np} - A_\infty \,|||_2 |||\, B_\infty \,|||_2, \\
a_2 &= m\beta |||\, I_{np} - A_\infty \,|||_2 |||\, B_\infty \,|||_2, \\
a_3 &= mh |||\, I_{np} - A_\infty \,|||_2, \\
a_4 &= n\beta g(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c), \\
a_5 &= h |||\, A_\infty \,|||_2, \\
a_6 &= g\sigma_B l\beta |||\, A - I_{np} \,|||_2, \\
a_7 &= g\sigma_B l\beta^2 |||\, B_\infty \,|||_2, \\
a_8 &= \sigma_B l\beta^2 |||\, B_\infty \,|||_2, \\
a_9 &= h\sigma_B l\beta.
\end{aligned}$$

When the step-size, η, satisfies

$$\eta < \min\left\{ \frac{\epsilon_1(1-\sigma_A)}{a_1\epsilon_1 + a_2\epsilon_2 + a_3\epsilon_3}, \frac{(1-\sigma_B)\epsilon_3 - \epsilon_1 a_6}{a_7\epsilon_1 + a_8\epsilon_2 + a_9\epsilon_3}, \frac{1}{n\beta\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c} \right\}, \tag{4.22}$$

where $\epsilon_1, \epsilon_2, \epsilon_3$ are positive constants such that

$$\epsilon_3 > 0, \qquad \epsilon_1 < \frac{(1-\sigma_B)\epsilon_3}{a_6}, \qquad \epsilon_2 > \frac{a_4\epsilon_1 + a_5\epsilon_3}{\alpha n(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c)}, \tag{4.23}$$

the spectral radius of $J(\eta)$, $\rho(J(\eta))$, is strictly less than 1, and therefore $\|\mathbf{x}(k) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2$ converges to zero geometrically at the rate of $O(\rho(J(\eta))^k)$.

Proof. Combining the results of Lemmas 15–18, one can verify that Eq. (4.19) holds if $\eta < \frac{2}{n\beta\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c}$. Recall that $\lambda = \max\left(\left|1 - \alpha n\eta(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c)\right|, \left|1 - \beta n\eta(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c)\right|\right)$. When $\eta < \frac{1}{n\beta\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c}$, $\lambda = 1 - \alpha n\eta(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c)$, since $\alpha \leq \beta$; see, e.g., Bubeck, 2014 for details. The goal is to find an upper bound of the step-size, $\widetilde{\eta}$, such that $\rho(J(\eta)) < 1$ when $\eta < \widetilde{\eta}$. In the light of Lemma 10, we solve for the range of the step-size, η, and a positive vector $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, \epsilon_3]^\top$ from the following linear matrix inequality (entry-wise):

$$\begin{bmatrix} \sigma_A + a_1\eta & a_2\eta & a_3\eta \\ a_4\eta & 1 - \alpha n\eta(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c) & a_5\eta \\ a_6 + a_7\eta & a_8\eta & \sigma_B + a_9\eta \end{bmatrix} \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix} < \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix}, \tag{4.24}$$

which is equivalent to the following set of inequalities:

$$\begin{cases} (a_1\epsilon_1 + a_2\epsilon_2 + a_3\epsilon_3)\eta & < & \epsilon_1(1-\sigma_A), \\ (a_4\epsilon_1 - \alpha n(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c)\epsilon_2 + a_5\epsilon_3)\eta & < & 0, \\ (a_7\epsilon_1 + a_8\epsilon_2 + a_9\epsilon_3)\eta & < & (1-\sigma_B)\epsilon_3 - \epsilon_1 a_6, \end{cases}$$

Solving the inequalities above, we have that when

$$\begin{cases} \epsilon_1 & < & \frac{(1-\sigma_B)\epsilon_3}{a_6}, \\ \epsilon_2 & > & \frac{a_4\epsilon_1 + a_5\epsilon_3}{\alpha n(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c)}, \\ \epsilon_3 & > & 0, \\ \eta & < & \min\left\{\frac{\epsilon_1(1-\sigma_A)}{a_1\epsilon_1 + a_2\epsilon_2 + a_3\epsilon_3}, \frac{(1-\sigma_B)\epsilon_3 - \epsilon_1 a_6}{a_7\epsilon_1 + a_8\epsilon_2 + a_9\epsilon_3}\right\}, \end{cases}$$

the inequality in Eq. (4.24) holds and the Theorem follows. □

4.4 Numerical Experiments

We consider a binary classification problem in the distributed setting, where we use logistic loss function to train a linear classifier. Each agent i has access to m_i training data, $(\mathbf{c}_{ij}, y_{ij}) \in \mathbb{R}^p \times \{-1, +1\}$, where $\mathbf{c}_{ij}$ contains p features of the jth training data at agent i and y_{ij} is the corresponding binary label. For privacy issues, agents do not share training data with each other. In order to use the entire data set for training, the network of agents cooperatively solves the following distributed logistic regression problem:

$$\min_{\mathbf{w}\in\mathbb{R}^p, b\in\mathbb{R}} F(\mathbf{w}, b) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} \ln\left[1 + \exp\left(-\left(\mathbf{w}^\top \mathbf{c}_{ij} + b\right) y_{ij}\right)\right] + \frac{\xi}{2}\|\mathbf{w}\|_2^2,$$

where the private function at each agent, i, is given by:

$$f_i(\mathbf{w}, b) = \sum_{j=1}^{m_i} \ln\left[1 + \exp\left(-\left(\mathbf{w}^\top \mathbf{c}_{ij} + b\right) y_{ij}\right)\right] + \frac{\xi}{2n}\|\mathbf{w}\|_2^2.$$

In our setting, $n = 8$, $p = 5$. The feature vectors, $\mathbf{c}_{ij}$'s, are Gaussian with zero mean and variance 2. The binary labels are randomly generated from standard Bernoulli distribution.

We compare the performance of the proposed algorithm with identical step-size in this paper, with ADD-OPT/Push-DIGing Xi, Xin, and Khan, 2017a; Nedić, Olshevsky, and Shi, 2017, FROST Xin, Xi, and Khan, 2018, and subgradient-push Tsianos, Lawlor, and Rabbat, 2012b; Nedić and Olshevsky, 2015, over the leftmost directed graph, $\mathcal{G}_1$, shown in Fig. 4.1. The simulation results are shown in the left figure in Fig. 4.2. Then we evaluate the proposed algorithm with identical step-size on the three different directed graphs, $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$,

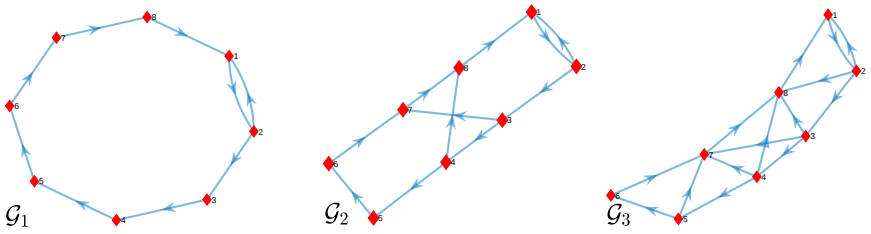

FIGURE 4.1: Strongly-connected but unbalanced directed graphs.

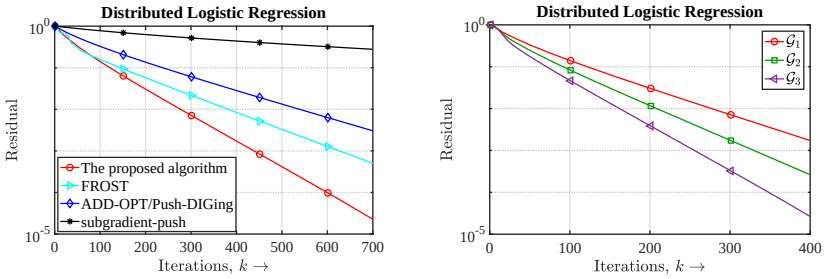

FIGURE 4.2: (Left) Comparison across different algorithms. (Right) Proposed algorithm over different graphs.we plot the average residuals at each agent, $\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i(k)-\mathbf{x}^*\|_2$.

shown in Fig 4.1, where each graph to the right has a few more edges compared to the one on its left. The simulation results are shown in the right figure in Fig 4.2. We note that the proposed linear algorithm achieves a comparable geometric (linear on the log-scale) convergence speed with other fast algorithms over directed graphs but with less computation and communication.

Chapter 5

Conclusion

In this thesis, we consider distributed optimization over directed graphs, where doubly-stochastic weight matrix cannot be constructed. Most of the existing algorithms are based on column-stochastic weights, which may be infeasible to implement in many practical scenarios. Row-stochastic weights, on the other hand, are straightforward to implement as each agent locally determines the weights. We propose a fast algorithm that we call FROST (Fast Row-stochastic Optimization with uncoordinated STep-sizes) and show that as long as the largest step-size is sufficiently small, FROST linearly converges to the global minimizer. Furthermore, we describe a completely linear distributed algorithm for optimization over directed graphs. The proposed linear algorithm achieves a comparable linear convergence rate with other fast methods over directed graphs yet with much less communication and computation as a result of avoiding push-sum (type) techniques. Our analysis is based on a novel approach where we establish simultaneous contractions of both row-and column-stochastic matrices under some arbitrary norms.

Bibliography

Benezit, F. et al. (2010). "Weighted Gossip: Distributed Averaging using non-doubly stochastic matrices". In: *IEEE International Symposium on Information Theory*, pp. 1753–1757. DOI: 10.1109/ISIT.2010.5513273.

Bubeck, S. (2014). "Convex optimization: Algorithms and complexity". In: *arXiv preprint arXiv:1405.4980*.

Cai, K. and H. Ishii (2012). "Average consensus on general strongly connected digraphs". In: *Automatica* 48.11, pp. 2750 –2761. ISSN: 0005-1098. DOI: http://dx.doi.org/10.1016/j.automatica.2012.08.003. URL: http://www.sciencedirect.com/science/article/pii/S0005109812004049.

Gharesifard, Bahman and Jorge Cortés (2012). "Distributed strategies for generating weight-balanced and doubly stochastic digraphs". In: *European Journal of Control* 18.6, pp. 539–557.

Horn, R. A. and C. R. Johnson (2013). *Matrix Analysis, 2nd ed.* New York, NY: Cambridge University Press.

Horn, Roger A and Charles R Johnson (1990). *Matrix analysis*. Cambridge university press.

Jakovetic, Dusan (2017). "A Unification, Generalization, and Acceleration of Exact Distributed First Order Methods". In: *arXiv preprint arXiv:1709.01317*.

Kempe, D., A. Dobra, and J. Gehrke (2003). "Gossip-based computation of aggregate information". In: *44th Annual IEEE Symposium on Foundations of Computer Science*, pp. 482–491. DOI: 10.1109/SFCS.2003.1238221.

Kushner, H. J. and G. Yin (2003). *Stochastic approximation and recursive algorithms and applications*. Vol. 35. Springer Science & Business Media.

Lee, S. and M. M. Zavlanos (2017). "Approximate Projection Methods for Decentralized Optimization with Functional Constraints". In: *IEEE Transactions on Automatic Control*.

Mai, Van Sy and Eyad H Abed (2016). "Distributed optimization over weighted directed graphs using row stochastic matrix". In: *American Control Conference (ACC), 2016*. IEEE, pp. 7165–7170.

Mansoori, F. and E. Wei (2017). "Superlinearly convergent asynchronous distributed network Newton method". In: *56th IEEE Annual Conference on Decision and Control*, pp. 2874–2879.

Mota, J. F. C. et al. (2012). "Distributed Basis Pursuit". In: *IEEE Transactions on Signal Processing* 60.4, pp. 1942–1956.

Nedić, A. and A. Olshevsky (2015). "Distributed optimization over time-varying directed graphs". In: *IEEE Trans. on Automatic Control* 60.3, pp. 601–615.

Nedić, A., A. Olshevsky, and W. Shi (2017). "Achieving Geometric Convergence for Distributed Optimization over Time-Varying Graphs". In: *SIAM Journal of Optimization*.

Nedić, A. and A. Ozdaglar (2009). "Distributed Subgradient Methods for Multi-Agent Optimization". In: *IEEE Trans. on Automatic Control* 54.1, pp. 48–61. ISSN: 0018-9286. DOI: 10.1109/TAC.2008.2009515.

Polyak, B. (1987). *Introduction to optimization*. Optimization Software.

Qu, G. and N. Li (2017a). "Accelerated distributed Nesterov gradient descent". In: *Arxiv: https://arxiv.org/abs/1705.07176*.

— (2017b). "Harnessing Smoothness to Accelerate Distributed Optimization". In: *IEEE Trans. on Control of Network Systems*.

— (2017c). "Harnessing smoothness to accelerate distributed optimization". In: *IEEE Transactions on Control of Network Systems*.

Raja, H. and W. U. Bajwa (2016). "Cloud K-SVD: A collaborative dictionary learing algorithm for big, distributed data". In: *IEEE Trans. Signal Processing* 64.1, pp. 173–188.

Safavi, Sam and Usman A Khan (2014). "On the convergence rate of swap-collide algorithm for simple task assignment". In: *Signals, Systems and Computers, 2014 48th Asilomar Conference on*. IEEE, pp. 1507–1510.

Shi, W. et al. (2015). "EXTRA: An Exact First-Order Algorithm for Decentralized Consensus Optimization". In: *SIAM Journal on Optimization* 25.2, pp. 944–966.

DOI: 10.1137/14096668X. eprint: http://dx.doi.org/10.1137/14096668X. URL: http://dx.doi.org/10.1137/14096668X.

Tsianos, K. I. (2013). "The role of the Network in Distributed Optimization Algorithms: Convergence Rates, Scalability, Communication/Computation Tradeoffs and Communication Delays". PhD thesis. Dept. Elect. Comp. Eng. McGill University.

Tsianos, K. I., S. Lawlor, and M. G. Rabbat (2012a). "Consensus-based distributed optimization: Practical issues and applications in large-scale machine learning". In: *50th Annual Allerton Conference on Communication, Control, and Computing*, pp. 1543–1550. DOI: 10.1109/Allerton.2012.6483403.

— (2012b). "Push-Sum Distributed Dual Averaging for convex optimization". In: *51st IEEE Annual Conference on Decision and Control*, pp. 5453–5458. DOI: 10.1109/CDC.2012.6426375.

Xi, C. and U. A. Khan (2016). "Distributed subgradient projection algorithm over directed graphs". In: *IEEE Trans. on Automatic Control* 62.8, pp. 3986–3992.

— (2017a). "DEXTRA: A fast algorithm for optimization over directed graphs". In: *IEEE Trans. on Automatic Control* 62.10, pp. 4980–4993.

Xi, C., Q. Wu, and U. A. Khan (2017). "On the distributed optimization over directed networks". In: *Neurocomputing* 267, pp. 508–515.

Xi, C., R. Xin, and U. A. Khan (2017a). "ADD-OPT: Accelerated Distributed Directed Optimization". In: *IEEE Trans. on Automatic Control. in press*.

Xi, C. et al. (2018). "Linear convergence in optimization over directed graphs with row-stochastic matrices". In: *IEEE Trans. on Automatic Control. in press*.

Xi, Chenguang and Usman A Khan (2017b). "Distributed subgradient projection algorithm over directed graphs". In: *IEEE Transactions on Automatic Control* 62.8, pp. 3986–3992.

Xi, Chenguang, Ran Xin, and Usman A Khan (2017b). "ADD-OPT: Accelerated distributed directed optimization". In: *IEEE Transactions on Automatic Control*.

Xin, R., C. Xi, and U. A. Khan (2018). "FROST – Fast row-stochastic optimization with uncoordinated step-sizes". In: *Arxiv: https://arxiv.org/abs/1803.09169*.

Xin, Ran and Usman A Khan (Mar. 4th, 2018). "A linear algorithm for optimization over directed graphs with geometric convergence". In: *ArXiv:* https://arxiv.org/abs/1803.02503.

Xu, J. et al. (2015). "Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes". In: *IEEE 54th Annual Conference on Decision and Control*, pp. 2055–2060.

Xu, Jinming et al. (2018). "Convergence of asynchronous distributed gradient methods over stochastic networks". In: *IEEE Transactions on Automatic Control* 63.2, pp. 434–448.

Ying, B. and A. H. Sayed (2018). "Performance limits of stochastic sub-gradient learning, part II: Multi-agent case". In: *Signal Processing* 144, pp. 253–264.

Yuan, K., Q. Ling, and W. Yin (2016). "On the convergence of decentralized gradient descent". In: *SIAM Journal on Optimization* 26.3, pp. 1835–1854.

Zhu, M. and S. Martínez (2010). "Discrete-time dynamic average consensus". In: *Automatica* 46.2, pp. 322–329.